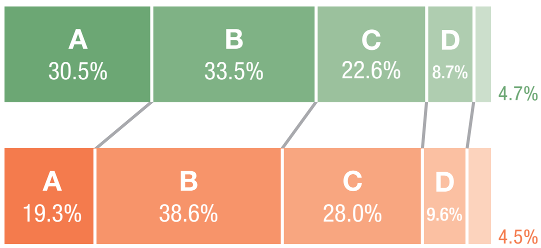
McCONNELL
BRUE
FLYNN

economics

LEARNSMART ADVANTAGE WORKS

LEARNSMART™



Without LearnSmart

More C students earn B's

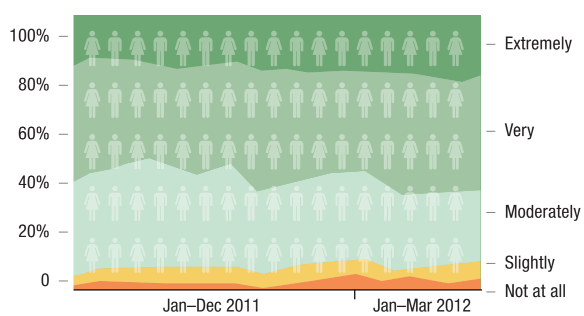
*Study: 690 students / 6 institutions

Over 20% more students pass the class with LearnSmart

*A&P Research Study

LEARNSMART™ Pass Rate - 70%

Without LearnSmart Pass Rate - 57%



More than 60% of all students agreed LearnSmart was a very or extremely helpful learning tool

*Based on 750,000 student survey responses

> AVAILABLE ON-THE-GO

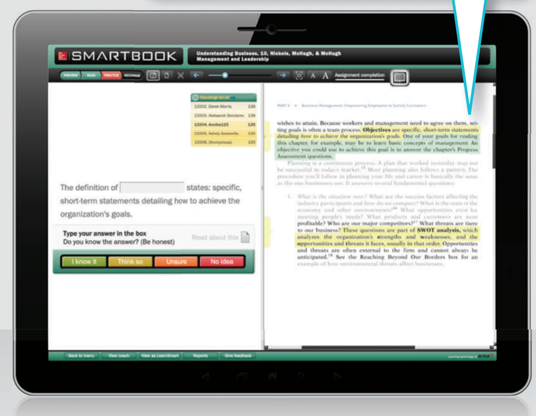
http://bit.ly/LS4Apple

http://bit.ly/LS4Droid

How do you rank against your peers?

Let's see how confident you are on the questions.

What you know (green) and what you still need to review (yellow), based on your answers.



COMPARE AND CHOOSE WHAT'S RIGHT FOR YOU

	BOOK	LEARNSMART	ASSIGNMENTS	
connect plus+	✓	✓	✓	LearnSmart, assignments, and SmartBook—all in one digital product for maximum savings!
connect plus+ Looseleaf	✓	✓	✓	Pop the pages into your own binder or carry just the pages you need.
connect plus+ Bound Book	✓	✓	✓	The #1 Student Choice!
SMARTBOOK™ Access Code	✓	✓		The first and only book that adapts to you!
LEARNSMART™ ADVANTAGE Access Code		✓		The smartest way to get from a B to an A.
CourseSmart eBook	✓			Save some green and some trees!
create™	✓	✓	✓	Check with your instructor about a custom option for your course.

> Buy directly from the source at www.ShopMcGraw-Hill.com.

Economics

PRINCIPLES, PROBLEMS, AND POLICIES

THE MCGRAW-HILL SERIES: ECONOMICS

ESSENTIALS OF ECONOMICS

Brue, McConnell, and Flynn
Essentials of Economics
Third Edition

Mandel
Economics: The Basics
Second Edition

Schiller
Essentials of Economics
Ninth Edition

PRINCIPLES OF ECONOMICS

Colander
Economics, Microeconomics, Macroeconomics
Ninth Edition

Frank and Bernanke
Principles of Economics, Principles of Microeconomics, Principles of Macroeconomics
Fifth Edition

Frank and Bernanke
Brief Editions: Principles of Economics, Principles of Microeconomics, Principles of Macroeconomics
Second Edition

Karlan and Morduch
Economics, Microeconomics, Macroeconomics
First Edition

McConnell, Brue, and Flynn
Economics, Microeconomics, Macroeconomics
Twentieth Edition

McConnell, Brue, and Flynn
Brief Editions: Microeconomics and Macroeconomics
Second Edition

Miller
Principles of Microeconomics
First Edition

Samuelson and Nordhaus
Economics, Microeconomics, Macroeconomics
Nineteenth Edition

Schiller
The Economy Today, The Micro Economy Today, The Macro Economy Today
Thirteenth Edition

Slavin
Economics, Microeconomics, Macroeconomics
Eleventh Edition

ECONOMICS OF SOCIAL ISSUES

Guell
Issues in Economics Today
Fifth Edition

Sharp, Register, and Grimes
Economics of Social Issues
Nineteenth Edition

ECONOMETRICS

Gujarati and Porter
Basic Econometrics
Fifth Edition

Gujarati and Porter
Essentials of Econometrics
Fourth Edition

MANAGERIAL ECONOMICS

Baye
Managerial Economics and Business Strategy
Eighth Edition

Brickley, Smith, and Zimmerman
Managerial Economics and Organizational Architecture
Fifth Edition

Thomas and Maurice
Managerial Economics
Tenth Edition

INTERMEDIATE ECONOMICS

Bernheim and Whinston
Microeconomics
Second Edition

Dornbusch, Fischer, and Startz
Macroeconomics
Eleventh Edition

Frank
Microeconomics and Behavior
Eighth Edition

ADVANCED ECONOMICS

Romer
Advanced Macroeconomics
Third Edition

MONEY AND BANKING

Cecchetti and Schoenholtz
Money, Banking, and Financial Markets
Third Edition

URBAN ECONOMICS

O'Sullivan
Urban Economics
Eighth Edition

LABOR ECONOMICS

Borjas
Labor Economics
Fifth Edition

McConnell, Brue, and Macpherson
Contemporary Labor Economics
Ninth Edition

PUBLIC FINANCE

Rosen and Gayer
Public Finance
Ninth Edition

Seidman
Public Finance
First Edition

ENVIRONMENTAL ECONOMICS

Field and Field
Environmental Economics: An Introduction
Fifth Edition

INTERNATIONAL ECONOMICS

Appleyard, Field, and Cobb
International Economics
Eighth Edition

King and King
International Economics, Globalization, and Policy: A Reader
Fifth Edition

Pugel
International Economics
Fifteenth Edition

THE SIX VERSIONS OF MCCONNELL, BRUE, FLYNN

Chapter*	*Economics*	*Microeconomics*	*Microeconomics: Brief Edition*	*Macroeconomics*	*Macroeconomics: Brief Edition*	*Essentials of Economics*
1. Limits, Alternatives, and Choices	x	x	x	x	x	x
2. The Market System and the Circular Flow	x	x	x	x	x	x
3. Demand, Supply, and Market Equilibrium	x	x	x	x	x	x
4. Market Failures: Public Goods and Externalities	x	x	**x**	x	x	**x**
5. Government's Role and Government Failure	x	x	**x**	x	**x**	**x**
6. Elasticity	x	x	**x**			**x**
7. Utility Maximization	x	x				
8. Behavioral Economics	x	x				
9. Businesses and the Costs of Production	x	x	**x**			x
10. Pure Competition in the Short Run	x	x	**x**			**x**
11. Pure Competition in the Long Run	x	x	**x**			**x**
12. Pure Monopoly	x	x	x			x
13. Monopolistic Competition and Oligopoly	x	x	x			x
13W. Technology, R&D, and Efficiency (Web Chapter)	x	x				
14. The Demand for Resources	x	x				
15. Wage Determination	x	x	x			x
16. Rent, Interest, and Profit	x	x				
17. Natural Resource and Energy Economics	x	x				
18. Public Finance: Expenditures and Taxes	x	x	**x**			
19. Antitrust Policy and Regulation	x	x				
20. Agriculture: Economics and Policy	x	x				
21. Income Inequality, Poverty, and Discrimination	x	x	**x**			**x**
22. Health Care	x	x				
23. Immigration	x	x				
24. An Introduction to Macroeconomics	x			x		
25. Measuring Domestic Output and National Income	x			x	**x**	**x**
26. Economic Growth	x			x	**x**	**x**
27. Business Cycles, Unemployment, and Inflation	x			x	x	x
28. Basic Macroeconomic Relationships	x			x		
29. The Aggregate Expenditures Model	x			x		
30. Aggregate Demand and Aggregate Supply	x			x	x	x
31. Fiscal Policy, Deficits, and Debt	x			x	x	x
32. Money, Banking, and Financial Institutions	x			x	x	x
33. Money Creation	x			x		
34. Interest Rates and Monetary Policy	x			x	x	x
35. Financial Economics	x			x		
36. Extending the Analysis of Aggregate Supply	x			x	**x**	
37. Current Issues in Macro Theory and Policy	x			x		
38. International Trade	x	x	**x**	x	**x**	**x**
39. The Balance of Payments, Exchange Rates, and Trade Deficits	x	x	**x**	x	**x**	**x**
39W. The Economics of Developing Countries (Web Chapter)	x			x		

*Chapter numbers refer to *Economics: Principles, Problems, and Policies.*
A red "**X**" indicates chapters that combine or consolidate content from two or more *Economics* chapters.

This page intentionally left blank

Twentieth Edition

Economics

PRINCIPLES, PROBLEMS, AND POLICIES

Campbell R. McConnell
University of Nebraska

Stanley L. Brue
Pacific Lutheran University

Sean M. Flynn
Scripps College

McGraw Hill Education

ECONOMICS: PRINCIPLES, PROBLEMS, AND POLICIES, TWENTIETH EDITION

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 DOW/DOW 1 0 9 8 7 6 5 4

ISBN 978-0-07-802175-6 (student edition)
MHID 0-07-802175-8 (student edition)
ISBN 978-0-07-766058-1 (instructor's edition)
MHID 0-07-766058-7 (instructor's edition)

Senior Vice President, Products & Markets: *Kurt L. Strand*
Vice President, Content Production & Technology Services: *Kimberly Meriwether David*
Managing Director: *Douglas Reiner*
Brand Manager: *Scott Smith*
Executive Director of Development: *Ann Torbert*
Development Editor: *Casey Rasch*
Director of Digital Content: *Doug Ruby*
Digital Development Editor: *Kevin Shanahan*
Marketing Manager: *Katie Hoenicke*
Director, Content Production: *Terri Schiesl*
Content Project Managers: *Bruce Gin, Lori Koetters*
Senior Buyer: *Michael R. McCormick*
Design: *Debra Kubiak*
Cover Image: *Ingram Publishing*
Lead Content Licensing Specialist: *Keri Johnson*
Typeface: *10/12 Janson Text Lt Std*
Compositor: *Aptara®, Inc.*
Printer: *R. R. Donnelley*

Library of Congress Cataloging-in-Publication Data

McConnell, Campbell R.
Economics : principles, problems, and policies / Campbell R. McConnell, University of Nebraska, Stanley L. Brue, Pacific Lutheran University, Sean M. Flynn, Scripps College.—Twentieth edition.
pages cm.—(The McGraw-Hill series in economics)
Includes index.
ISBN 978-0-07-802175-6 (student edition : alk. paper)—ISBN 0-07-802175-8 (student edition : alk. paper)—ISBN 978-0-07-766058-1 (instructor's edition : alk. paper)—ISBN 0-07-766058-7 (instructor's edition : alk. paper)
1. Economics. I. Brue, Stanley L., 1945- II. Flynn, Sean Masaki. III. Title.
HB171.5.M47 2015
330—dc23

2013038889

www.mhhe.com

To Mem and to Terri and Craig, and to past instructors

ABOUT THE AUTHORS

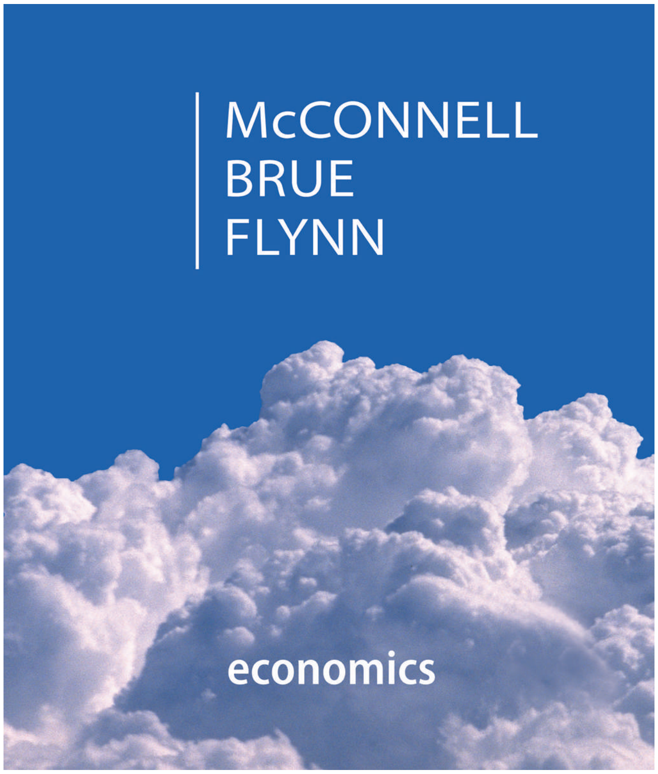
CAMPBELL R. MCCONNELL earned his Ph.D. from the University of Iowa after receiving degrees from Cornell College and the University of Illinois. He taught at the University of Nebraska–Lincoln from 1953 until his retirement in 1990. He is also coauthor of *Contemporary Labor Economics*, ninth edition; *Essentials of Economics*, second edition; *Macroeconomics: Brief Edition;* and *Microeconomics: Brief Edition* (all The McGraw-Hill Companies), and has edited readers for the principles and labor economics courses. He is a recipient of both the University of Nebraska Distinguished Teaching Award and the James A. Lake Academic Freedom Award and is past president of the Midwest Economics Association. Professor McConnell was awarded an honorary Doctor of Laws degree from Cornell College in 1973 and received its Distinguished Achievement Award in 1994. His primary areas of interest are labor economics and economic education. He has an extensive collection of jazz recordings and enjoys reading jazz history.

STANLEY L. BRUE did his undergraduate work at Augustana College (South Dakota) and received its Distinguished Achievement Award in 1991. He received his Ph.D. from the University of Nebraska–Lincoln. He is retired from a long career at Pacific Lutheran University, where he was honored as a recipient of the Burlington Northern Faculty Achievement Award. Professor Brue has also received the national Leavey Award for excellence in economic education. He has served as national president and chair of the Board of Trustees of Omicron Delta Epsilon International Economics Honorary. He is coauthor of *Economic Scenes*, fifth edition (Prentice-Hall); *Contemporary Labor Economics*, ninth edition; *Essentials of Economics*, second edition; *Macroeconomics: Brief Edition; Microeconomics: Brief Edition* (all The McGraw-Hill Companies); and *The Evolution of Economic Thought*, seventh edition (South-Western). For relaxation, he enjoys international travel, attending sporting events, and skiing with family and friends.

SEAN M. FLYNN did his undergraduate work at the University of Southern California before completing his Ph.D. at U.C. Berkeley, where he served as the Head Graduate Student Instructor for the Department of Economics after receiving the Outstanding Graduate Student Instructor Award. He teaches at Scripps College (of the Claremont Colleges) and is the author of *Economics for Dummies* (Wiley) and coauthor of *Essentials of Economics*, second edition; *Macroeconomics: Brief Edition;* and *Microeconomics: Brief Edition* (all The McGraw-Hill Companies). His research interests include finance, behavioral economics, and health economics. An accomplished martial artist, he has represented the United States in international aikido tournaments and is the author of *Understanding Shodokan Aikido* (Shodokan Press). Other hobbies include running, traveling, and enjoying ethnic food.

LIST OF KEY GRAPHS

PREFACE



Welcome to the 20th edition of *Economics*, the best-selling economics textbook in the world. An estimated 15 million students have used *Economics* or its companion editions, *Macroeconomics* and *Microeconomics*. *Economics* has been adapted into Australian and Canadian editions and translated into Italian, Russian, Chinese, French, Spanish, Portuguese, and other languages. We are pleased that *Economics* continues to meet the market test: nearly one out of five U.S. students in principles courses used the 19th edition.

Fundamental Objectives

We have three main goals for *Economics:*

- Help the beginning student master the principles essential for understanding the economizing problem, specific economic issues, and policy alternatives.
- Help the student understand and apply the economic perspective and reason accurately and objectively about economic matters.
- Promote a lasting student interest in economics and the economy.

Student Feedback

The twentieth edition has a renewed focus on today's students and their various approaches to learning. How do today's students study? How are they using mobile technology? When are they using the textbook, and how are they using the textbook? To help answer these questions, McGraw-Hill and author Sean Flynn formed a Student Advisory Board consisting of students from Belmont University, the University of Louisiana–Lafayette, Tarrant County College, and West Virginia University Institute of Technology. The Student Advisory Board participated in a wide variety of evaluation and testing activities over six months and provided targeted recommendations to improve the 20th edition and its ancillary learning materials. Their feedback was incredibly valuable, and the authors incorporated their suggestions in this revision.

What's New and Improved?

One of the benefits of writing a successful text is the opportunity to revise—to delete the outdated and install the new, to rewrite misleading or ambiguous statements, to introduce more relevant illustrations, to improve the organizational structure, and to enhance the learning aids.

We trust that you will agree that we have used this opportunity wisely and fully. Some of the more significant changes include the following.

Restructured Introductory Chapters

We have divided the five-chapter grouping of introductory chapters common to *Economics*, *Microeconomics*, and *Macroeconomics* into two parts. Part 1 contains Chapter 1 (Limits, Alternatives, and Choices) and Chapter 2 (The Market System and the Circular Flow). The content in Part 2 has changed and now consists of the following three chapters: Chapter 3 (Demand, Supply, and Market Equilibrium), Chapter 4 (Market Failures: Public Goods and Externalities), and Chapter 5 (Government's Role and Government Failure).

As restructured, the three chapters that now form Part 2 give students a panorama of:

- The efficiency and allocation benefits of competitive markets.
- How and why governments can help when there are cases of market failure.
- An appreciation of government failure so that students do not assume that government intervention is an easy or guaranteed panacea for the misallocations and inefficiencies caused by market failure.

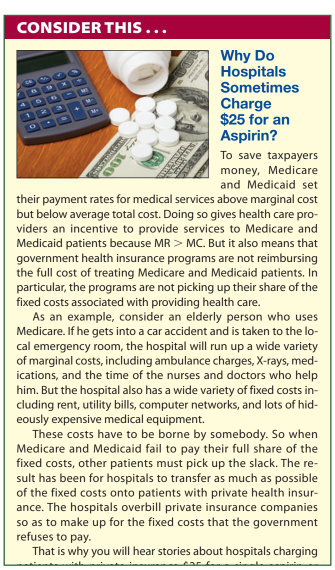
Our new approach responds to suggestions by reviewers to:

- Move the elasticity chapter back into *Microeconomics*.
- Boost the analysis of government failure to help students better understand many of the problems currently besetting the U.S. economy.

Our new approach embraces these suggestions. For microeconomics instructors, the new ordering provides a clear supply-and-demand path to the subsequent chapters on consumer and producer behavior while also giving students a stronger policy background on not only market failures but government interventions in the economy and whether they are likely to improve efficiency. For macroeconomics instructors, the new sequence provides a theoretical grounding that can help students better understand issues such as excessive government deficit spending and why there may be insufficient regulation of the financial sector. And because Chapters 4 and 5 are both optional and modular, instructors can skip them if they wish to move directly from Chapter 3's discussion of supply and demand to the core microeconomics or macroeconomics chapters.

New "Consider This" and "Last Word" Pieces

Our "Consider This" boxes are used to provide analogies, examples, or stories that help drive home central economic ideas in a student-oriented, real-world manner.



CONSIDER THIS . . .

Why Do Hospitals Sometimes Charge $25 for an Aspirin?

To save taxpayers money, Medicare and Medicaid set their payment rates for medical services above marginal cost but below average total cost. Doing so gives health care providers an incentive to provide services to Medicare and Medicaid patients because MR > MC. But it also means that government health insurance programs are not reimbursing the full cost of treating Medicare and Medicaid patients. In particular, the programs are not picking up their share of the fixed costs associated with providing health care.

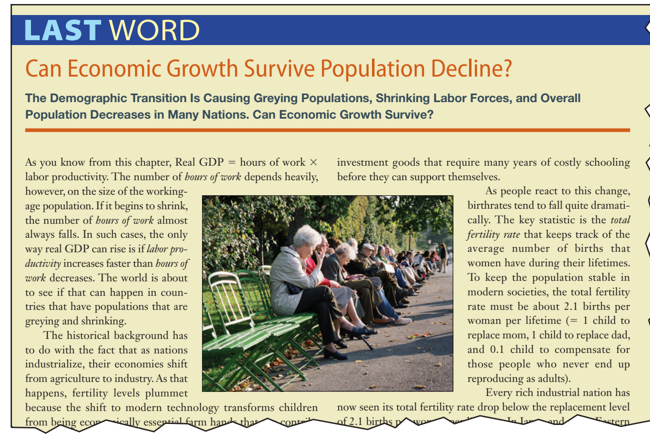
As an example, consider an elderly person who uses Medicare. If he gets into a car accident and is taken to the local emergency room, the hospital will run up a wide variety of marginal costs, including ambulance charges, X-rays, medications, and the time of the nurses and doctors who help him. But the hospital also has a wide variety of fixed costs including rent, utility bills, computer networks, and lots of hideously expensive medical equipment.

These costs have to be borne by somebody. So when Medicare and Medicaid fail to pay their full share of the fixed costs, other patients must pick up the slack. The result has been for hospitals to transfer as much as possible of the fixed costs onto patients with private health insurance. The hospitals overbill private insurance companies so as to make up for the fixed costs that the government refuses to pay.

That is why you will hear stories about hospitals charging

For instance, a "Consider This" box titled "McHits and McMisses" illustrates consumer sovereignty through a listing of successful and unsuccessful products. How businesses exploit price discrimination is driven home in a "Consider This" box that explains why ballparks charge different admission prices for adults and children but only one set of prices at their concession stands. These brief vignettes, each accompanied by a photo, illustrate key points in a lively, colorful, and easy-to-remember way. We have added 14 new "Consider This" boxes in this edition.

Our "Last Word" pieces are lengthier applications or case studies that are placed near the end of each chapter.



LAST WORD

Can Economic Growth Survive Population Decline?

The Demographic Transition Is Causing Greying Populations, Shrinking Labor Forces, and Overall Population Decreases in Many Nations. Can Economic Growth Survive?

As you know from this chapter, Real GDP = hours of work × labor productivity. The number of *hours of work* depends heavily, however, on the size of the working-age population. If it begins to shrink, the number of *hours of work* almost always falls. In such cases, the only way real GDP can rise is if *labor productivity* increases faster than *hours of work* decreases. The world is about to see if that can happen in countries that have populations that are greying and shrinking.

The historical background has to do with the fact that as nations industrialize, their economies shift from agriculture to industry. As that happens, fertility levels plummet because the shift to modern technology transforms children

investment goods that require many years of costly schooling before they can support themselves.

As people react to this change, birthrates tend to fall quite dramatically. The key statistic is the *total fertility rate* that keeps track of the average number of births that women have during their lifetimes. To keep the population stable in modern societies, the total fertility rate must be about 2.1 births per woman per lifetime (= 1 child to replace mom, 1 child to replace dad, and 0.1 child to compensate for those people who never end up reproducing as adults).

Every rich industrial nation has now seen its total fertility rate drop below the replacement level

For example, the "Last Word" section for Chapter 1 (Limits, Alternatives, and Choices) examines pitfalls to sound economic reasoning, while the "Last Word" section for Chapter 4 (Market Failures: Public Goods and Externalities) examines cap-and-trade versus carbon taxes as policy responses to excessive carbon dioxide emissions. There are 14 new "Last Word" sections in this edition.

If you are unfamiliar with *Economics*, we encourage you to thumb through the chapters to take a quick look at these highly visible features.

New Chapter on Government's Role and Government Failure

We have responded to instructor suggestions by placing this new chapter on Government's Role and Government Failure into the introductory section of the book. Its early placement gives students a taste of political economy and the practical difficulties with government regulation and intervention. Topics covered include the special-interest effect, rent seeking, regulatory capture, political corruption, unfunded liabilities, and unintended consequences.

The chapter begins, however, by reminding students of government's great power to improve equity and efficiency. When read along with Chapter 4 on market failure, this new chapter on government failure should provide students with a balanced perspective. After learning why government intervention is needed to counter market failures, they will also learn that governments often have difficulty in fulfilling their full potential for improving economic outcomes.

An optional appendix incorporates the material on public choice theory and voting paradoxes that was formerly located in Chapter 17 of the 19th edition. Instructors wishing to give their students an even deeper appreciation of government failure may wish to assign this material.

Meanwhile, the material on asymmetric information that was located in Chapter 17 of the 19th edition has been moved into an appendix attached to the current edition's Chapter 4 on market failure. That way, instructors wishing to give their students a deeper look at market failure will have the material on asymmetric information located immediately after that chapter's discussion of public goods and externalities.

New Chapter on Behavioral Economics

By building upon the material on prospect theory that appeared in Chapter 6 of the 19th edition, we have created a new full-length chapter on behavioral economics for the 20th edition. Topics covered include time inconsistency, myopia, decision-making heuristics, framing effects, mental accounting, loss aversion, the endowment effect, and reciprocity. The discussion is couched in terms of consumer decision making and includes numerous concrete examples to bring the material home for students.

We have also striven to make clear to students the ways in which behavioral economics builds upon and augments the insights of traditional neoclassical economics. Thus, the chapter opens with a section comparing and contrasting behavioral economics and neoclassical economics so that students will be able to see how both can be used in tandem to help understand and predict human choice behavior.

The chapter is designed, however, to be modular. So instructors may skip it completely without any fear that its concepts are needed to understand subsequent chapters.

New Discussions of the Financial Crisis and the Recession

Our modernization of the macroeconomics in the 18th edition has met with great success, measured by reviews, instructor feedback, and market response. We recast the entire macro analysis in terms of the modern, dominant paradigm of macroeconomics, using economic growth as the central backdrop and viewing business fluctuations as significant and costly variations in the rate of growth. In this paradigm, business cycles result from demand shocks (or, less often, supply shocks) in conjunction with inflexible short-run product prices and wages. The degree of price and wage stickiness decreases with time. In our models, the *immediate short run* is a period in which both the price level and wages are not only sticky, but stuck; the *short run* is a period in which product prices are flexible but wages are not; and the *long run* is a period in which both product prices and wages are fully flexible. Each of these three periods—and thus each of the models based on them—is relevant to understanding the actual macro economy and its occasional difficulties.

In this edition, we have mainly focused on incorporating into our new macroeconomic schema an analysis of the financial crisis, the recession, and the hesitant recovery. We first introduce the debate over the policy response to the 2007–2009 recession in Chapter 24 (An Introduction to Macroeconomics) via a new "Last Word" that briefly lays out the major opposing viewpoints about the nature and size of the stimulus that was applied during and after the crisis. In Chapter 25 (Measuring Domestic Output and National Income), we point out that the main flows in the National Income and Product Accounts usually expand over time, but not always, as demonstrated by the recession. In Chapter 26 (Economic Growth), we discuss how the recession relates to the growth/production possibilities dynamics of Figure 26.2. In Chapter 27 (Business Cycles, Unemployment, and Inflation), we have a new "Last Word" that discusses the very slow recovery in employment after the Great Recession.

In Chapter 28 (Basic Macroeconomic Relationships), we include two "Consider This" boxes, one on how the paradox of thrift applied to consumer behavior during the recession and the other on the riddle of plunging investment spending at the same time the interest rate dropped to near zero during the recession. In Chapter 29 (The Aggregate Expenditures Model), we use the recession as a timely application of how a decline in aggregate expenditures can produce a recessionary expenditure gap and a highly negative GDP gap. Chapter 30 (Aggregate Demand and Aggregate Supply) features a new "Last Word" on the debate among economists as to why the recovery from the 2007–2009 recession was so slow despite the historically unprecedented amounts of monetary and fiscal stimulus that were applied by policymakers. Chapter 31 (Fiscal Policy, Deficits, and Debt) provided a terrific opportunity to bring each of these timely and relevant subjects up-to-date, and we took full advantage of that opportunity.

In Chapter 32 (Money, Banking, and Financial Institutions), we updated the major section on the financial crisis and added a new "Last Word" on whether the existence of megabanks that are considered "too big to fail" has led prosecutors to hold off on the full enforcement of securities laws and banking regulations. Chapter 33 (Money Creation) includes a new "Last Word" on the potential dangers of excessive leverage in the financial system and whether, consequently, regulators should increase required reserve ratios.

Chapter 34 (Interest Rates and Monetary Policy) features several new discussions relating to Fed policies during the recession and recovery, including quantitative easing, the zero interest rate policy, and Operation Twist. While giving the Fed high marks for dealing with the crisis and its aftermath, we also point out that some economists think the Fed contributed to the financial crisis by keeping interest rates too low for too long during the recovery from the 2001 recession. Chapter 35 (Financial Economics) presented a new opportunity for us to demonstrate how a sharp decline of the "appetite for risk" alters the slope of the Security Market Line (SML) and changes investment patterns between stocks and bonds.

Other mentions of the recession and subsequent recovery are spread throughout the remainder of the macro chapters, including in the discussions of macro debates, trade protectionism, and trade deficits. Although we found these various ways to work the recession and recovery into our macro chapters, we are confident that our basic macroeconomic models will serve equally well in explaining expansion back to the economy's historical growth path. The new inclusions simply help students see the relevance of the models to what they are seeing in the news and perhaps experiencing in their own lives. The overall tone of the book, including the macro, continues to be optimistic with respect to the long-term growth prospects of market economies.

Reorganized and Extended End-of-Chapter Questions and Problems

The 19th edition featured separate sections for end-of-chapter Questions and Problems. Due to strong demand on the part of instructors for an increase in the number of problems that are both autogradable and algorithmic, we have for the 20th edition added about 10 new problems per chapter and have, in addition, revised our organizational scheme for questions and problems.

The questions and problems are now divided into three categories: Discussion Questions, Review Questions, and Problems.

- The Discussion Questions are analytic and often allow for free responses.
- The Review Questions focus on the apprehension of key concepts but are worded so as to always require specific answers, thereby allowing for autograding and algorithmic variation.
- The Problems are quantitative and require specific answers so that they, too, are both autogradable and algorithmic.

All of the questions and problems are assignable through McGraw-Hill's *Connect Economics* and we have additionally aligned all of the questions and problems with the learning objectives presented at the beginning of chapters. The new structure as well as the newly added problems were well received by reviewers, many of them long-time users of the book.

Current Discussions and Examples

The 20th edition of *Economics* refers to and discusses many current topics. Examples include surpluses and shortages of tickets at the Olympics; the myriad impacts of ethanol subsidies; creative destruction; applications of behavioral economics; applications of game theory; the most rapidly expanding and disappearing U.S. jobs; oil and gasoline prices; cap-and-trade systems and carbon taxes; the value-added tax; state lotteries; consumption versus income inequality; the impact of electronic medical records on health care costs; the surprising fall in illegal immigration after the 2007–2009 recession; the massive increase in long-term unemployment; the difficulty of targeting fiscal stimulus; the rapid rise in college tuition; the slow recovery from the Great Recession; ballooning federal budget deficits and public debt; the long-run funding shortfalls in Social Security and Medicare; the effect of rising dependency ratios on economic growth; innovative Federal Reserve policies including quantitative easing, the zero interest rate policy, and explicit inflation targets; the massive excess reserves in the banking system; the jump in the size of the Fed's balance sheet; the effect of the zero interest rate policy on savers; regulation of "too big to fail" banks; trade adjustment assistance; the European Union and the eurozone; changes in exchange rates; and many other current topics.

Chapter-by-Chapter Changes

Each chapter of *Economics*, 20th edition, contains updated data reflecting the current economy, revised Learning Objectives, and reorganized and expanded end-of-chapter content. Several chapters also contain one or more additional Quick Review boxes to help students review and solidify content as they are reading along.

Chapter-specific updates include:

Chapter 1: Limits Alternatives, and Choices features three refreshed "Consider This" pieces, a more concise definition of macroeconomics, and wording improvements that clarify the main concepts.

Chapter 2: The Market System and the Circular Flow contains a heavily revised introductory section on the different types of economic systems found in the world today as well as a new section on how the market system deals with risk and uncertainty. Reviewers asked for more material on risk and its effects on economic behavior. This short section provides a brief, nontechnical framework for students to understand how the market economy deals with risk and uncertainty. There is also a new "Consider This" box on how insurance encourages investment by transferring risk from those who do not wish to bear it to those who are willing to bear it as a business proposition.

Chapter 3: Demand, Supply, and Market Equilibrium contains a short new section in the appendix that introduces students to markets with vertical supply curves so that the concept of perfectly inelastic supply will come more easily to microeconomics students and the concept of vertical long-run aggregate supply will come more easily to macroeconomics students.

Chapter 4: Market Failures: Public Goods and Externalities includes a new "Consider This" piece on how musicians have reacted to the reality that Internet file sharing has transformed recorded music from a private good into a public good. There is also a new appendix that explains market failures caused by asymmetric information. The appendix gives instructors the option of extending and deepening this chapter's study of market failure. Its content previously appeared in Chapter 17 of the 19th edition.

Chapter 5: Government's Role and Government Failure is a new chapter that offers a balanced treatment of both the great benefits as well as the possible drawbacks of government economic intervention and regulation. The chapter includes topics of interest for both microeconomics and macroeconomics students, such as: regulatory capture, unfunded liabilities, the collective-action problem, bureaucratic inertia, the tendency for politicians to run budget deficits to please voters, and the special-interest effect. So while Chapter 4 makes the case for government regulation to compensate for market failures, this new chapter introduces students to the fact that government interventions are themselves susceptible to both allocative and productive inefficiency. As noted previously, the chapter also includes an appendix that incorporates the voting and public choice material that appeared in Chapter 17 of the 19th edition for instructors who wish to present their students with the most prominent theoretical models dealing with government failure.

Chapter 6: Elasticity contains a new "Consider This" vignette that relates elasticity to the high cost of college tuition as well as a number of wording improvements to increase clarity. This chapter was previously located in the 19th edition directly after Chapter 3 on supply and demand. It has been move to this new location to serve as the first chapter of the new three-chapter Part 3 that deals with consumer behavior. Along those lines, this chapter on elasticity explains in greater depth than Chapter 3 how consumers and producers react to changes in prices.

Chapter 7: Utility Maximization contains the utility-maximization material that previously appeared as the first half of the 19th edition's Chapter 6 (Consumer Behavior). This core content has been refreshed with a new example using iPads to explain consumer equilibrium.

Chapter 8: Behavioral Economics is a new, chapter-length overview of behavioral economics. The chapter incorporates the short section on prospect theory that was located in Chapter 6 of the 19th edition. New concepts include a discussion of the human brain's cognitive limitations and dependence on heuristics, how time inconsistency and myopia cause people to make suboptimal long-run decisions, and how people's sense of fairness and reciprocity affects decision making. The discussion of prospect theory includes anchoring, mental accounting, loss aversion, and the endowment effect.

Chapter 9: Businesses and the Costs of Production has a new "Consider This" vignette on sunk costs as well as a new "Last Word" on how additive manufacturing and 3-D printing may replace mass production with mass customization.

Chapter 10: Pure Competition in the Short Run is mostly unchanged from the 19th edition (where it appeared as Chapter 8). It contains several wording changes to improve clarity as well as a new Quick Review to increase retention.

Chapter 11: Pure Competition in the Long Run features a new "Last Word" that discusses the possibility that in certain industries patent protections may hinder rather than help innovation and the process of creative destruction.

Chapter 12: Pure Monopoly has a new "Last Word" on how network effects and economies of scale have driven the monopolistic growth of Internet giants such as Facebook, Google, and Amazon. We have also revised our explanation of barriers to entry in monopoly industries with high fixed costs. The revised presentation builds upon the new material in Chapter 2 that covers risk and its effects on economic decision making.

Chapter 13: Monopolistic Competition and Oligopoly contains several updated examples as well as a new "Last Word" on the intense oligopolistic competition that has ensued between major Internet companies like Google, Apple, and Microsoft as they have attempted to compete in each other's core lines of business.

Chapter 13 Web: Technology, R&D, and Efficiency contains a new "Last Word" discussing the drastic decline in federal research and development (R&D) spending over the past 50 years.

Chapter 14: The Demand for Resources features extensive data updates and a new Quick Review summarizing the material toward the end of the chapter.

Chapter 15: Wage Determination contains a new "Consider This" box on fringe benefits. It makes the point that if workers in a competitive labor market want higher fringe benefits, they must accept lower take-home pay (because total compensation is fixed by the equilibrium wage).

Chapter 16: Rent, Interest, and Profit features extensive data updating and several new examples in addition to additional Quick Reviews and nearly a dozen new Review Questions.

Chapter 17: Natural Resource and Energy Economics contains numerous data updates, an additional Quick Review, and two new "Consider This" vignettes. The first reports on alternative energy subsidies and the surge in oil and gas production resulting from hydraulic fracking. The second introduces students to Garret Hardin's "The Tragedy of the Commons" story as a way of understanding most resource-depletion crises.

Chapter 18: Public Finance: Expenditures and Taxes features numerous data updates, a new Quick Review, and a reference to the recent study that found that the United States has the most progressive tax system among all 30 OECD nations.

Chapter 19: Antitrust Policy and Regulation has many new examples of real-world collusive behavior and an expanded explanation of cartel behavior in the section on legal cartel theory.

Chapter 20: Agriculture: Economics and Policy features numerous data updates as well as a new discussion of the recent run-up in agricultural commodity prices.

Chapter 21: Income Inequality, Poverty, and Discrimination contains major data updates, enhanced figure captions, and several short additions that add clarity regarding the specific arguments made in the debate over income inequality.

Chapter 22: Health Care features three new "Consider This" pieces. The first explains how Medicare's decision to reimburse medical services at rates above marginal cost but below total cost forces hospitals to transfer the difference onto those with private health insurance. The second relates how electronic medical records have unexpectedly lowered care and increased costs. The third describes some of the problems being encountered with the implementation of the Patient Protection and Affordable Care Act (Obamacare).

Chapter 23: Immigration contains extensive data updates as well as a new "Last Word" piece covering the startling decline in illegal immigration that happened during and after the Great Recession of 2007–2009.

Chapter 24: An Introduction to Macroeconomics benefits from extensive revisions to the chapter's header structure, several new Quick Reviews, and a new "Last Word" that covers in a brief and accessible form the major opposing policy viewpoints about the effectiveness and ideal size of government stimulus during and after the 2007–2009 recession.

Chapter 25: Measuring Domestic Output and National Income features two clarifications driven directly by student input. First, the table giving U.S. GDP by both the expenditure method and the income method is more clearly referenced in all instances so as to reduce any possible confusion as to which part of the table is being referred to. Second, there is now a more detailed explanation of the statistical discrepancy that appears when the income method is used to calculate GDP.

Chapter 26: Economic Growth benefits from extensive data updates, a new Quick Review to help solidify comprehension, and a new "Last Word" that discusses the challenges to economic growth posed by falling birth rates and a greying population.

Chapter 27: Business Cycles, Unemployment, and Inflation features a new "Last Word" on the slow recovery of employment after the Great Recession. There are also two new "Consider This" vignettes that discuss, respectively, the relationship between downwardly sticky wages and unemployment and the idea that moderate inflation rates may help to lower unemployment by allowing firms to cut real wages without cutting nominal wages.

Chapter 28: Basic Macroeconomic Relationships features a revised header structure to better guide students through the material, a new Quick Review to help solidify retention, and substantial revisions to several graphs and their captions to further refine and clarify the fundamental concepts introduced in this chapter.

Chapter 29: The Aggregate Expenditures Model has substantial changes to a key figure in order to improve clarity as well as a heavily revised list of Learning Objectives.

Chapter 30: Aggregate Demand and Aggregate Supply features a new "Last Word" on the discussion economists have been having as to why the recovery from the Great Recession has been so slow despite record amounts of monetary and fiscal stimulus.

Chapter 31: Fiscal Policy, Deficits, and Debt features extensive data updates to help students understand the historically unprecedented size of recent federal budget deficits.

Chapter 32: Money, Banking, and Financial Institutions features a new "Last Word" on how some banks are now considered "too big to fail" and how that has affected the prosecution of financial crimes. The chapter also contains four new Quick Reviews to help students better retain the chapter's material.

Chapter 33: Money Creation features a new "Last Word" on the dangers of leverage in the banking system and whether required reserve ratios should consequently be increased.

Chapter 34: Interest Rates and Monetary Policy features a new section on the Fed's monetary policy initiatives after the Great Recession, including quantitative easing (QE), forward guidance, the zero interest rate policy (ZIRP), and Operation Twist. There is also a new "Last Word" discussing the potential unintended consequences of QE and ZIRP.

Chapter 35: Financial Economics features extensive data updates, five new Quick Reviews, and revised section headers to increase clarity.

Chapter 36: Extending the Analysis of Aggregate Supply features data updates, an extended discussion of the Laffer Curve, and a new Quick Review.

Chapter 37: Current Issues in Macro Theory and Policy has a new Quick Review plus a brief discussion of the Fed's recent decisions to (1) have an explicit inflation target and (2) preannounce the likely duration of open-market operations and quantitative easing.

Chapter 38: International Trade features extensive data updates, several revised figure captions, and three new Quick Reviews.

Chapter 39: The Balance of Payments, Exchange Rates, and Trade Deficits features revised problems, extensive data updates, four new Quick Reviews, and a new set of Learning Objectives.

Chapter 39 Web: The Economics of Developing Countries features extensive data revisions and an all-new set of Discussion Questions.

COI1: The United States in the Global Economy features a heavily revised map of the international distribution of income levels, a totally new set of Discussion Questions, and extensive data updates.

COI2: Previous International Exchange Rate Systems remains unchanged for the 20th edition.

Distinguishing Features

Comprehensive Explanations at an Appropriate Level *Economics* is comprehensive, analytical, and challenging yet fully accessible to a wide range of students. The thoroughness and accessibility enable instructors to select topics for special classroom emphasis with confidence that students can read and comprehend other independently assigned material in the book. Where needed, an extra sentence of explanation is provided. Brevity at the expense of clarity is false economy.

Fundamentals of the Market System Many economies throughout the world are still making difficult transitions from planning to markets while a handful of other countries such as Venezuela seem to be trying to reestablish government-controlled, centrally planned economies. Our detailed description of the institutions and operation of the market system in Chapter 2 (The Market System and the Circular Flow) is therefore even more relevant than before. We pay particular attention to property rights, entrepreneurship, freedom of enterprise and choice, competition, and the role of profits because these concepts are often misunderstood by beginning students worldwide.

Extensive Treatment of International Economics We give the principles and institutions of the global economy extensive treatment. The appendix to Chapter 3 (Demand, Supply, and Market Equilibrium) has an application on exchange rates. Chapter 38 (International Trade) examines key facts of international trade, specialization and comparative advantage, arguments for protectionism, impacts of tariffs and subsidies, and various trade agreements. Chapter 39 (The Balance of Payments, Exchange Rates, and Trade Deficits) discusses the balance of payments, fixed and floating exchange rates, and U.S. trade deficits. Web Chapter 39 (The Economics of Developing Countries) takes a look at the special problems faced by developing countries and how the advanced industrial countries try to help them.

As noted previously in this preface, Chapter 38 (International Trade) is constructed such that instructors who want to cover international trade early in the course can assign it immediately after Chapter 3. Chapter 38 requires only a good understanding of production possibilities analysis and supply and demand analysis to comprehend. International competition, trade flows, and financial flows are integrated throughout the micro and macro sections. "Global Perspective" boxes add to the international flavor of the book.

Early and Extensive Treatment of Government The public sector is an integral component of modern capitalism. This book introduces the role of government early. Chapter 4 (Market Failures: Public Goods and Externalities) systematically discusses public goods and government policies toward externalities. Chapter 5 (Government's Role and Government Failure) details the factors that cause government failure. And Chapter 18 (Public Finance: Expenditures and Taxes) examines taxation and government expenditures in detail. Both the micro and the macro sections of the text include issue- and policy-oriented chapters.

Stress on the Theory of the Firm We have given much attention to microeconomics in general and to the theory of the firm in particular, for two reasons. First, the concepts of microeconomics are difficult for most beginning students; abbreviated expositions usually compound these difficulties by raising more questions than they answer. Second, we wanted to couple analysis of the various market structures with a discussion of the impact of each market arrangement on price, output levels, resource allocation, and the rate of technological advance.

Step-by-Step, Two-Path Macro As in the previous edition, our macro continues to be distinguished by a systematic step-by-step approach to developing ideas and building models. Explicit assumptions about price and wage stickiness are posited and then systematically peeled away, yielding new models and extensions, all in the broader context of growth, expectations, shocks, and degrees of price and wage stickiness over time.

In crafting this step-by-step macro approach, we took care to preserve the "two-path macro" that many instructors appreciated. Instructors who want to bypass the immediate short-run model (Chapter 29: The Aggregate Expenditures Model) can proceed without loss of continuity directly to the short-run AD-AS model (Chapter 30: Aggregate Demand and Aggregate Supply), fiscal policy, money and banking, monetary policy, and the long-run AD-AS analysis.

Emphasis on Technological Change and Economic Growth This edition continues to emphasize economic growth. Chapter 1 (Limits, Alternatives, and Choices) uses the production possibilities curve to show the basic ingredients of growth. Chapter 26 (Economic Growth) explains how growth is measured and presents the facts of growth. It also discusses the causes of growth, looks at productivity growth, and addresses some controversies surrounding economic growth. Chapter 26's "Last Word" examines whether economic growth can survive demographic decline. Web Chapter 39 focuses on developing countries and the growth obstacles they confront. Web Chapter 13 (Technology, R&D, and Efficiency) provides an explicit and cohesive discussion of the microeconomics of technological advance, including topics such as invention, innovation, and diffusion; start-up firms; R&D decision making; market structure and R&D effort; and creative destruction.

Focus on Economic Policy and Issues For many students, the micro chapters on antitrust, agriculture, income inequality, health care, and immigration, along with the macro chapters on fiscal policy and monetary policy, are where the action is centered. We guide that action along logical lines through the application of appropriate analytical tools. In the micro, we favor inclusiveness; instructors can effectively choose two or three chapters from Part 6.

Integrated Text and Web Site *Economics* and its Web site are highly integrated through in-text Web buttons, bonus Web chapters, multiple-choice self-tests at the Web site, math notes, and other features. Our Web site is part and parcel of our student learning package, customized to the book.

The in-text Web buttons (or indicators) merit special mention. Two differently colored rectangular indicators appear throughout the book, informing readers that complementary content on a subject can be found at our Web site, **www.mcconnell20e.com.** The indicator types are:

WORKED PROBLEMS

W1.1

Budget lines

Worked Problems Written by Norris Peterson of Pacific Lutheran University (WA), these pieces consist of side-by-side computational questions and computational procedures used to derive the answers. In essence, they extend the textbook's explanations of various computations—for example, of real GDP, real GDP per capita, the unemployment rate, the inflation rate, per-unit production costs, economic profit, and more. From a student's perspective, they provide "cookbook" help for solving numerical problems.

ORIGIN OF THE IDEA

O1.1

Origin of the term "economics"

Origin of the Ideas These pieces, written by Randy Grant of Linfield College (OR), are brief histories of 70 major ideas discussed in the book. They identify the particular economists who developed ideas such as opportunity cost, equilibrium price, the multiplier, comparative advantage, and elasticity.

Organizational Alternatives

Although instructors generally agree on the content of principles of economics courses, they sometimes differ on how to arrange the material. *Economics* includes 11 parts, and thus provides considerable organizational flexibility. We place microeconomics before macroeconomics because this ordering is consistent with how contemporary economists view the direction of linkage between the two components. The introductory material of Parts 1 and 2, however, can be followed immediately by the macroanalysis of Parts 7 and 8. Similarly, the two-path macro enables covering the full aggregate expenditures model or advancing directly from the basic macro relationships chapter to the AD-AS model.

Some instructors will prefer to intersperse the microeconomics of Parts 4 and 5 with the problems chapters of

Part 6. Chapter 20 on agriculture may follow Chapters 10 and 11 on pure competition; Chapter 19 on antitrust and regulation may follow Chapters 12, 13, and 13Web on imperfect competition models and technological advance. Chapter 23 on immigration may follow Chapter 15 on wages; and Chapter 21 on income inequality may follow Chapters 15 and 16 on distributive shares of national income.

Instructors who teach the typical two-semester course and feel comfortable with the book's organization will find that, by putting Parts 1 to 6 in the first semester and Parts 7 to 11 in the second, the material is divided logically between the two semesters.

Finally, Chapter 38 on international trade can easily be moved up to immediately after Chapter 3 on supply and demand for instructors who want an early discussion of international trade.

Pedagogical Aids

Economics is highly student-oriented. The "To the Student" statement at the beginning of Part 1 details the book's many pedagogical aids. The 20th edition is also accompanied by a variety of high-quality supplements that help students master the subject and help instructors implement customized courses.

Supplements for Students and Instructors

Study Guide One of the world's leading experts on economic education, William Walstad of the University of Nebraska–Lincoln, prepared the *Study Guide*. Many students find either the printed or digital version indispensable. Each chapter contains an introductory statement, a checklist of behavioral objectives, an outline, a list of important terms, fill-in questions, problems and projects, objective questions, and discussion questions.

The *Guide* comprises a superb "portable tutor" for the principles student. Separate *Study Guides* are available for the macro and micro paperback editions of the text.

Instructor's Manual Shawn Knabb of Western Washington University revised and updated the *Instructor's Manuals* to accompany the 20th edition of the text. The revised *Instructor's Manual* includes:

- Chapter summaries.
- Listings of "what's new" in each chapter.
- Teaching tips and suggestions.
- Learning objectives.
- Chapter outlines.
- Extra questions and problems.
- Answers to the end-of-chapter questions and problems, plus correlation guides mapping content to the learning objectives.

The *Instructor's Manual* is available on the instructor's side of the Online Learning Center.

Three Test Banks Test Bank I contains about 6,500 multiple-choice and true-false questions, most of which were written by the text authors. Randy Grant revised Test Bank I for the 20th edition. Test Bank II contains around 6,000 multiple-choice and true-false questions, updated by Felix Kwan of Maryville University. All Test Bank I and II questions are organized by learning objective, topic, AACSB Assurance of Learning, and Bloom's Taxonomy guidelines. Test Bank III, written by William Walstad, contains more than 600 pages of short-answer questions and problems created in the style of the book's end-of-chapter questions. Test Bank III can be used to construct student assignments or design essay and problem exams. Suggested answers to the essay and problem questions are included. In all, more than 14,000 questions give instructors maximum testing flexibility while ensuring the fullest possible text correlation.

Test Banks I and II are available in *Connect Economics*, through EZ Test Online, and in MS Word. EZ Test allows professors to create customized tests that contain both questions that they select from the test banks as well as questions that they craft themselves. Test Bank III is available in MS Word on the password-protected instructor's side of the Online Learning Center, and on the Instructor Resource CD.

PowerPoint Presentations The PowerPoint Presentations for the 20th edition were updated by a dedicated team of instructors: Stephanie Campbell of Mineral Area College, Amy Chataginer of Mississippi Gulf Coast Community College, and Shannon Aucoin of the University of Louisiana at Lafayette. Each chapter is accompanied by a concise yet thorough tour of the key concepts. Instructors can use these Web site presentations in the classroom, and students can use them on their computers.

Digital Image Library Every graph and table in the text is available on the instructor's side of the Web site and on the Instructor's Resource CD-ROM.

Scanning Barcodes For students using smartphones and tablets, scanning barcodes (or QR codes) located within the chapter guide students to additional chapter resources, including:

- Web buttons
- Student PowerPoints
- Worked problems

Try It Now!

Students not using smartphones or tablets can access the same resources by clicking the barcodes when viewing the eBook or by going to **www.mcconnell20e.com.**

Online Learning Center (www.mcconnell20e.com) The Web site accompanying this book is a central resource for students and instructors alike. The optional Web Chapters (Chapter 13W: Technology, R&D, and Efficiency and Chapter 39W: The Economics of Developing Countries) plus the two Content Options for Instructors (The United States in the Global Economy and Previous International Exchange-Rate Systems), are posted as full-color PDF files. The in-text Web buttons alert the students to points in the book where they can springboard to the Web site to get more information. Students can also review PowerPoint presentations and test their knowledge of a chapter's concepts with a self-graded multiple-choice quiz. The password-protected Instructor Center houses the Instructor's Manual, all three Test Banks, and links to EZ Test Online, PowerPoint presentations, and the Digital Image Library.

Computerized Test Bank Online A comprehensive bank of test questions is provided within McGraw-Hill's flexible electronic testing program EZ Test Online (**www.eztestonline.com**). EZ Test Online allows instructors to simply and quickly create tests or quizzes for their students. Instructors can select questions from multiple McGraw-Hill test banks or author their own, and then either print the finalized test or quiz for paper distribution or publish it online for access via the Internet.

This user-friendly program allows instructors to sort questions by format; select questions by learning objectives or Bloom's taxonomy tags; edit existing questions or add new ones; and scramble questions for multiple versions of the same test. Instructors can export their tests for use in WebCT, Blackboard, and PageOut, making it easy to share assessment materials with colleagues, adjuncts, and TAs. Instant scoring and feedback are provided, and EZ Test Online's record book is designed to easily export to instructor gradebooks.

Assurance-of-Learning Ready Many educational institutions are focused on the notion of assurance of learning, an important element of some accreditation standards. *Economics* is designed to support your assurance-of-learning initiatives with a simple yet powerful solution. Each chapter in the book begins with a list of numbered learning objectives to which each end-of-chapter question and problem is then mapped. In this way, student responses to those questions and problems can be used to assess how well students are mastering each particular learning objective. Each test bank question for *Economics* also maps to a specific learning objective.

You can use our test bank software, EZ Test Online, or *Connect Economics* to easily query for learning outcomes and objectives that directly relate to the learning objectives for your course. You can then use the reporting features to aggregate student results in a similar fashion, making the collection and presentation of assurance-of-learning data simple and easy.

AACSB Statement The McGraw-Hill Companies is a proud corporate member of AACSB International. Understanding the importance and value of AACSB accreditation, *Economics*, 20th edition, has sought to recognize the curricula guidelines detailed in the AACSB standards for business accreditation by connecting end-of-chapter questions in *Economics*, 20th edition, and the accompanying test banks to the general knowledge and skill guidelines found in the AACSB standards.

This AACSB Statement for *Economics*, 20th edition, is provided only as a guide for the users of this text. The AACSB leaves content coverage and assessment within the purview of individual schools, their respective missions, and their respective faculty. While *Economics*, 20th edition, and the teaching package make no claim of any specific AACSB qualification or evaluation, we have, within *Economics*, 20th edition, labeled selected questions according to the six general knowledge and skills areas.

Digital Solutions

McGraw-Hill *Connect® Economics*

Less Managing. More Teaching. Greater Learning. *Connect Economics* is an online assignment and assessment solution that offers a number of powerful tools and features that make managing assignments easier so faculty can spend more time teaching. With *Connect Economics*, students can engage with their coursework anytime and anywhere, making the learning process more accessible and efficient.

Simple Assignment Management With *Connect Economics*, creating assignments is easier than ever, so you can spend more time teaching and less time managing. The assignment management function enables you to:

- Create and deliver assignments easily with selectable end-of-chapter questions and test bank items.
- Streamline lesson planning, student progress reporting, and assignment grading to make classroom management more efficient than ever.

- Go paperless with online submission and grading of student assignments.

Smart Grading *Connect Economics* helps students learn more efficiently by providing feedback and practice material when they need it, where they need it. The grading function enables instructors to:

- Score assignments automatically, giving students immediate feedback on their work and side-by-side comparisons with correct answers.
- Access and review each response; manually change grades or leave comments for students to review.
- Reinforce classroom concepts with practice tests and instant quizzes.

Instructor Library The *Connect Economics* Instructor Library is your repository for additional resources to improve student engagement in and out of class. You can select and use any asset that enhances your lecture.

Student Study Center The *Connect Economics* Student Study Center is the place for students to access additional resources. The Student Study Center offers students quick access to lectures, practice materials, eBooks, study questions, and more.

Student Progress Tracking *Connect Economics* keeps instructors informed about how each student, section, and class is performing, allowing for more productive use of lecture and office hours. The progress-tracking function enables instructors to:

- View scored work immediately and track individual or group performance with assignment and grade reports.
- Access a real-time view of student or class performance relative to learning objectives.
- Collect data and generate reports required by many accreditation organizations like AACSB.

McGraw-Hill *Connect® Plus Economics*

McGraw-Hill reinvents the textbook learning experience for the modern student with *Connect Plus Economics*. A seamless integration of an eBook and *Connect Economics*, *Connect Plus Economics* provides all of the *Connect Economics* features plus the following:

- An integrated eBook, allowing for anytime, anywhere access to the textbook.
- Dynamic links between the problems or questions you assign to your students and the location in the eBook where that problem or question is covered.
- A powerful search function to pinpoint and connect key concepts in a snap.

In short, *Connect Plus Economics* offers you and your students powerful tools and features that optimize your time and energies, enabling you to focus on course content, teaching, and student learning. *Connect Plus Economics* also offers a wealth of content resources for both instructors and students. This state-of-the-art, thoroughly tested system supports you in preparing students for the world that awaits.

For more information about *Connect*, go to **www.mcgrawhillconnect.com**, or contact your local McGraw-Hill sales representative.

LearnSmart Advantage

New from McGraw-Hill Education, LearnSmart Advantage is a series of adaptive learning products fueled by LearnSmart, the most widely used and intelligent adaptive learning resource on the market. Developed to deliver demonstrable results in boosting grades, increasing course retention, and strengthening memory recall, LearnSmart Advantage spans the entire learning process, from course preparation to the first adaptive reading experience. A smarter learning experience for students coupled with valuable reporting tools for instructors, LearnSmart Advantage is advancing learning like no other product in higher education today. Products in McConnell's LearnSmart Advantage Suite include:

LearnSmart LearnSmart is one of the most effective and successful adaptive learning resources in the market today, proven to strengthen memory recall, keep students in class, and boost grades. Distinguishing what students know from what they don't, and honing in on the concepts they are most likely to forget, LearnSmart continuously adapts to each student's needs to build an individual learning path so students study smarter and retain more knowledge. Reports provide valuable insight to instructors, so precious class time can be spent on higher-level concepts and discussion.

LearnSmart Achieve LearnSmart Achieve is a revolutionary new learning system that combines a continually adaptive learning experience with necessary course resources to focus students on mastering concepts they don't already know. The program adjusts to each student individually as he or she progresses, creating just-in-time learning experiences by presenting interactive content that is tailored to each student's needs. A convenient time-management feature and reports for instructors also ensure students stay on track.

SmartBook SmartBook is the first and only adaptive reading experience available today. SmartBook changes reading from a passive and linear experience to an engaging and dynamic one in which students are more likely to master and retain important concepts, coming to class better prepared. Valuable reports provide instructors insight as to how students are progressing through textbook content, and are useful for shaping in-class time or assessment.

This revolutionary technology suite is available only from McGraw-Hill Education. To learn more, go to **http://learnsmartadvantage.com** or contact your representative for a demo.

Tegrity Campus: Lectures 24/7

Tegrity Campus is a service that makes class time available 24/7 by automatically capturing every lecture in a searchable format for students to review when they study and complete assignments. With a simple one-click start-and-stop process, you capture all computer screens and corresponding audio. Students can replay any part of any class with easy-to-use browser-based viewing on a PC or Mac.

Educators know that the more students can see, hear, and experience class resources, the better they learn. In fact, studies prove it. With Tegrity Campus, students quickly recall key moments by using Tegrity Campus's unique search feature. This search function helps students efficiently find what they need, when they need it, across an entire semester of class recordings. Help turn all your students' study time into learning moments immediately supported by your lecture.

To learn more about Tegrity, you can watch a two-minute Flash demo at **tegritycampus.mhhe.com.**

CourseSmart

CourseSmart is a new way for faculty to find and review eTextbooks. It's also a great option for students who are interested in accessing their course materials digitally. CourseSmart offers thousands of the most commonly adopted textbooks across hundreds of courses from a wide variety of higher education publishers. It is the only place for faculty to review and compare the full text of a textbook online. At CourseSmart, students can save up to 50% off the cost of a print book, reduce their impact on the environment, and gain access to powerful Web tools for learning, including full text search, notes and highlighting, and e-mail tools for sharing notes between classmates. Complete tech support is also included with each title. Finding your eBook is easy. Visit **www.CourseSmart.com** and search by title, author, or ISBN.

McGraw-Hill Customer Care Contact Information

Getting the most from new technology can be challenging. So McGraw-Hill offers a large suite of complementary support services for faculty using *Economics*. You can contact our Product Specialists 24 hours a day to set up online technology instruction. Or you can contact customer support at any time by either calling **800-331-5094** or by visiting the Customer Experience Group (CXG) Support Center at **www.mhhe.com/support.** They will put you in touch with a Technical Support Analyst familiar with *Economics* and its technology offerings. And, of course, our online knowledge bank of Frequently Asked Questions is always available at the just-mentioned Web site for instant answers to the most common technology questions.

Acknowledgments

We give special thanks to Norris Peterson and Randy Grant, who created the "button" content on our Web site. We again thank James Reese of the University of South Carolina at Spartanburg, who wrote the original Internet exercises. Although many of those questions were replaced or modified in the typical course of revision, several remain virtually unchanged. Ryan Umbeck, Peter Staples, and Heather Schumacker deserve considerable credit for their hard work on the questions and problems in *Connect*. Thanks to the many dedicated instructors who accuracy-checked the end-of-chapter content, test banks, and Instructor's Manuals: Jennifer Pate, Charles Harrington, Rick Hirschi, Melissa Rueterbusch, Gregory McGiffney, and Mike Winterhalter. Finally, we thank William Walstad and Tom Barbiero (the coauthor of our Canadian edition) for their helpful ideas and insights.

We are greatly indebted to an all-star group of professionals at McGraw-Hill—in particular Douglas Reiner, Lori Koetters, Casey Rasch, Scott Smith, Bruce Gin, and Katie Hoenicke—for their publishing and marketing expertise. We thank Keri Johnson for her selection of the "Consider This" and "Last Word" photos and Debra Kubiak for the design.

The 20th edition has benefited from a number of perceptive formal reviews. The reviewers, listed at the end of the preface, were a rich source of suggestions for this revision. To each of you, and others we may have inadvertently overlooked, thank you for your considerable help in improving *Economics*.

Sean M. Flynn
Stanley L. Brue
Campbell R. McConnell

REVIEWERS

Richard Agesa, *Marshall University*
Yamin Ahmad, *University of Wisconsin–Whitewater*
Eun Ahn, *University of Hawaii, West Oahu*
Thomas Andrews, *West Chester University of Pennsylvania*
Fatma Antar, *Manchester Community College*
Len Anyanwu, *Union County College*
Emmanuel Asigbee, *Kirkwood Community College*
John Atkins, *Pensacola State College*
Moses Ayiku, *Essex County College*
Wendy Bailey, *Troy University*
Dean Baim, *Pepperdine University*
Herman Baine, *Broward College*
Tyra Barrett, *Pellissippi State Community College*
Jill Beccaris-Pescatore, *Montgomery County Community College*
Kevin Beckwith, *Salem State University*
Christian Beer, *Cape Fear Community College*
Robert Belsterling, *Pennsylvania State University, Altoona*
Laura Jean Bhadra, *Northern Virginia Community College, Manassas*
Priscilla Block, *Broward College*
Augustine Boakye, *Essex County College*
Stephanie Campbell, *Mineral Area College*
Bruce Carpenter, *Mansfield University*
Tom Cate, *Northern Kentucky University*
Semih Emre Çekin, *Texas Tech University*
Suparna Chakraborty, *University of San Francisco*
Claude Chang, *Johnson & Wales University*
Amy Chataginer, *Mississippi Gulf Coast Community College–Gautier*
Shuo Chen, *State University of New York–Geneseo*
Jon Chesbro, *Montana Tech of the University of Montana*
Amod Choudhary, *Lehman College*
Constantinos Christofides, *East Stroudsburg University*
Jane Clary, *College of Charleston*
Jane Cline, *Forsyth Technical Community College*
Anthony Daniele, *St. Petersburg College–Gibbs*
Rosa Lee Danielson, *College of DuPage*
Ribhi Daoud, *Sinclair Community College*
William L. Davis, *University of Tennessee–Martin*
Richard Dixon, *Thomas Nelson Community College*
Tanya Downing, *Cuesta College*
Scott Dressler, *Villanova University*
Brad Duerson, *Des Moines Area Community College*
Mark J. Eschenfelder, *Robert Morris University*
Maxwell Eseonu, *Virginia State University*
Michael Fenick, *Broward College*
Tyrone Ferdnance, *Hampton University*
Jeffrey Forrest, *St. Louis Community College–Florissant Valley*
Mark Frascatore, *Clarkson University*
Shelby Frost, *Georgia State University*
Sudip Ghosh, *Penn State University–Berks*
Daniel Giedeman, *Grand Valley State University*
Scott Gilbert, *Southern Illinois University*
James Giordano, *Villanova University*
Susan Glanz, *St. John's University*

Lowell Glenn, *Utah Valley University*
Terri Gonzales, *Delgado Community College*
Moonsu Han, *North Shore Community College*
Virden Harrison, *Modesto Junior College*
Richard R. Hawkins, *University of West Florida*
Kim Hawtrey, *Hope College*
Glenn Haynes, *Western Illinois University*
Michael Heslop, *Northern Virginia Community College*, *Annandale*
Jesse Hoyt Hill, *Tarrant County College*
Calvin Hoy, *County College of Morris*
Jim Hubert, *Seattle Central Community College*
Greg W. Hunter, *California State Polytechnic University*, *Pomona*
Christos Ioannou, *University of Minnesota–Minneapolis*
Faridul Islam, *Utah Valley University*
Mahshid Jalilvand, *University of Wisconsin–Stout*
Ricot Jean, *Valencia Community College–Osceola*
Jonatan Jelen, *City College of New York*
Brad Kamp, *University of South Florida*, *Sarasota-Manatee*
Kevin Kelley, *Northwest Vista College*
Chris Klein, *Middle Tennessee State University*
Barry Kotlove, *Edmonds Community College*
Richard Kramer, *New England College*
Felix Kwan, *Maryville University*
Ted Labay, *Bishop State Community College*
Tina Lance, *Germanna Community College–Fredericksburg*
Yu-Feng Lee, *New Mexico State University–Las Cruces*
Adam Y.C. Lei, *Midwestern State University*
Phillip Letting, *Harrisburg Area Community College*
Brian Lynch, *Lake Land College*
Zagros Madjd-Sadjadi, *Winston-Salem State University*
Laura Maghoney, *Solano Community College*
Vincent Mangum, *Grambling State University*
Benjamin Matta, *New Mexico State University–Las Cruces*
Pete Mavrokordatos, *Tarrant County College–Northeast Campus*
Frederick May, *Trident Technical College*
Katherine McClain, *University of Georgia*
Michael McIntyre, *Copiah-Lincoln Community College*
Robert McKizzie, *Tarrant County College–Southeast Campus*
Kevin McWoodson, *Moraine Valley Community College*
Edwin Mensah, *University of North Carolina at Pembroke*
Randy Methenitis, *Richland College*
Ida Mirzaie, *The Ohio State University*
David Mitch, *University of Maryland–Baltimore County*
Ramesh Mohan, *Bryant University*
Daniel Morvey, *Piedmont Technical College*
Shahriar Mostashari, *Campbell University*
Ted Muzio, *St. John's University*
Cliff Nowell, *Weber State University*
Albert Okunade, *University of Memphis*
Mary Ellen Overbay, *Seton Hall University*
Tammy Parker, *University of Louisiana at Monroe*
Alberto Alexander Perez, *Harford Community College*
David Petersen, *American River College*
Mary Anne Pettit, *Southern Illinois University–Edwardsville*
Jeff Phillips, *Morrisville State College*
Robert Poulton, *Graceland University*
Dezzie Prewitt, *Rio Hondo College*
Joe Prinzinger, *Lynchburg College*
Jaishankar Raman, *Valparaiso University*
Natalie Reaves, *Rowan University*
Virginia Reilly, *Ocean County College*
Tim Reynolds, *Alvin Community College*
John Romps, *Saint Anselm College*
Tom Scheiding, *Elizabethtown College*
Amy Schmidt, *Saint Anselm College*
Ron Schuelke, *Santa Rosa Junior College*
Alexandra Shiu, *McLennan Community College*
Dorothy Siden, *Salem State University*
Timothy Simpson, *Central New Mexico Community College*
Jonathan Sleeper, *Indian River State College*
Jose Rodriguez Solis, *Northern Virginia Community College*
Camille Soltau-Nelson, *Oregon State University*
Robert Sonora, *Fort Lewis College*
Nick Spangenberg, *Ozarks Technical Community College*
Dennis Spector, *Naugatuck Valley Community College*
Thomas Stevens, *University of Massachusetts*, *Amherst*
Tamika Steward, *Tarrant County College*, *Southeast*
Robin Sturik, *Cuyahoga Community College Western–Parma*
Travis Taylor, *Christopher Newport University*
Ross Thomas, *Central New Mexico Community College*
Mark Thompson, *Augusta State University*
Deborah Thorsen, *Palm Beach State College*
Michael Toma, *Armstrong Atlantic State University*
Dosse Toulaboe, *Fort Hays State University*
Jeff Vance, *Sinclair Community College*
Cheryl Wachenheim, *North Dakota State University–Fargo*
Christine Wathen, *Middlesex County College*
Wendy Wysocki, *Monroe County Community College*
Edward Zajicek, *Winston-Salem State University*

BRIEF CONTENTS

CONTENTS

This page intentionally left blank

PART **ONE**

INTRODUCTION TO ECONOMICS AND THE ECONOMY

TO THE STUDENT

Economics students learn more, understand more, and make better choices for themselves and others. This book and its interactive learning materials like LearnSmart will speed you on your way to mastering the principles of economics and making a better life for yourself and those around you.

This book and its ancillaries contain several features designed to help you learn economics:

- ***Web buttons (indicators)*** A glance through the book reveals many pages with rectangular icons set into the text. These "buttons" alert you to helpful learning aids available with the book. The orange button symbolizes "Worked Problems." Numeric problems are presented and then solved, side-by-side, step-by-step. Seeing how the problems are worked will help you solve similar problems on quizzes and exams. The green button stands for "Origin of the Idea." Each of these pieces traces a particular economic idea to the people who first developed it.

  After reading a chapter, look back at the chapter's Web buttons and their associated numbers. On the home page of our Internet site, **www.mcconnell20e.com**, select Student Edition and use the pull-down list under "Choose one" to find the Web button content for each chapter. You can also scan the accompanying QR code with your smartphone or tablet. Need a barcode reader? Try ScanLife, available in your app store.

WORKED PROBLEMS

W2.1
Least-cost production

ORIGIN OF THE IDEA

O1.4
Ceteris paribus

- ***Other Internet aids*** Our Web site contains many other aids. In the Student Edition you will find self-testing multiple-choice quizzes, PowerPoint presentations, and much more. For those of you with a very strong mathematics background, be sure to note the "See the Math" section on the Web site. There you will find nearly 50 notes that develop the algebra and, in a few cases, the calculus that underlie the economic concepts.
- ***Appendix on graphs*** Be assured, however, that you will need only basic math skills to do well in your introductory economics course. In particular, you will need to be comfortable with basic graphical analysis and a few quantitative concepts. The appendix at the end of Chapter 1 reviews graphs and slopes of curves. You may want to read it before starting Chapter 1.
- ***Reviews*** Each chapter contains several Quick Reviews as well as an end-of-chapter summary. These review sections will help you focus on essential ideas and study for exams.
- ***Key terms and Key Graphs*** Key terms are set in boldface type within the chapters, listed at the end of each chapter, and again defined in the glossary at the end of the book. Graphs with special importance are labeled Key Graphs, and each includes a multiple-choice Quick Quiz. Your instructor may or may not emphasize all of these figures, but you should pay special attention to those that are discussed in class; you can be certain there will be exam questions on them.
- ***Consider This and Last Word boxes*** Many chapters include a "Consider This" box. These brief pieces provide commonplace analogies, examples, and stories that help you understand and remember central economic ideas. Each chapter concludes with a "Last Word" box. Some of them are revealing applications of economic ideas; others are short case studies. While it is tempting to ignore in-text boxes, don't. Most are fun to read, and all will improve your grasp of economics.
- ***Questions and Problems*** The end of each chapter features separate sections of Discussion Questions, Review Questions, and Problems. The Discussion Questions are analytic and often ask for free responses, and the Review Questions require either specific answers to short computations or brief definitions about important concepts. The Problems involve longer computations for which a series of specific answers must be given. Each is keyed to a particular learning objective (LO) in the list of LOs at the beginning of the chapter. At the Web site is a multiple-choice quiz for each chapter.
- ***Study Guide*** We enthusiastically recommend the *Study Guide* accompanying this text. This "portable tutor" contains not only a broad sampling of various kinds of questions but a host of useful learning aids. Software-driven tutorials, including the Self Quiz and Study in *Connect Economics*, are also available with the text.

Our two main goals are to help you understand and apply economics and help you improve your analytical skills. An understanding of economics will enable you to comprehend a whole range of economic, social, and political problems that otherwise would seem puzzling and perplexing. Also, your study will enhance reasoning skills that are highly prized in the workplace.

Good luck with your study. We think it will be well worth your time and effort.

CHAPTER **1**

Limits, Alternatives, and Choices

Learning Objectives

LO1.1 **Define economics and the features of the economic perspective.**

LO1.2 **Describe the role of economic theory in economics.**

LO1.3 **Distinguish microeconomics from macroeconomics and positive economics from normative economics.**

LO1.4 **Explain the individual's economizing problem and how trade-offs, opportunity costs, and attainable combinations can be illustrated with budget lines.**

LO1.5 **List the categories of scarce resources and delineate the nature of society's economizing problem.**

LO1.6 **Apply production possibilities analysis, increasing opportunity costs, and economic growth.**

LO1.7 **Explain how economic growth and international trade increase consumption possibilities.**

LO1.8 **(Appendix) Understand graphs, curves, and slopes as they relate to economics.**

(An appendix on understanding graphs follows this chapter. If you need a quick review of this mathematical tool, you might benefit by reading the appendix first.) People's wants are numerous and varied. Biologically, people need only air, water, food, clothing, and shelter. But in modern societies people also desire goods and services that provide a more comfortable or affluent standard of living. We want bottled water, soft drinks, and fruit juices, not just water from the creek. We want salads, burgers, and pizzas, not just berries and nuts. We want jeans, suits, and coats, not just woven reeds. We want apartments, condominiums, or houses, not just mud huts. And, as the saying goes, "That is not

the half of it." We also want flat-panel TVs, Internet service, education, national defense, cell phones, health care, and much more.

Fortunately, society possesses productive resources, such as labor and managerial talent, tools and machinery, and land and mineral deposits. These resources, employed in the economic system (or simply the economy), help us produce goods and services that satisfy many of our economic wants. But the blunt reality is that our economic wants far exceed the productive capacity of our scarce (limited) resources. We are forced to make choices. This unyielding truth underlies the definition of **economics,** which is the social science concerned with how individuals, institutions, and society make optimal (best) choices under conditions of scarcity.

ORIGIN OF THE IDEA

O1.1

Origin of the term "Economics"

The Economic Perspective

LO1.1 Define economics and the features of the economic perspective.

Economists view things from a unique perspective. This **economic perspective,** or economic way of thinking, has several critical and closely interrelated features.

Scarcity and Choice

The economic resources needed to make goods and services are in limited supply. This **scarcity** restricts options and demands choices. Because we "can't have it all," we must decide what we will have and what we must forgo.

At the core of economics is the idea that "there is no free lunch." You may be treated to lunch, making it "free" from your perspective, but someone bears a cost. Because all resources are either privately or collectively owned by members of society, ultimately society bears the cost. Scarce inputs of land, equipment, farm labor, the labor of cooks and waiters, and managerial talent are required. Because society could have used these resources to produce something else, it sacrifices those other goods and services in making the lunch available. Economists call such sacrifices **opportunity costs:** To obtain more of one thing, society forgoes the opportunity of getting the next best thing. That sacrifice is the opportunity cost of the choice.

Purposeful Behavior

Economics assumes that human behavior reflects "rational self-interest." Individuals look for and pursue opportunities to increase their **utility**—the pleasure, happiness, or satisfaction obtained from consuming a good or service. They allocate their time, energy, and money to maximize their

CONSIDER THIS . . .

Free for All?

Free products are seemingly everywhere. Sellers offer free apps, free cell phones, and free checking accounts. Dentists give out free toothbrushes. At state visitor centers, there are free brochures and maps.

Does the presence of so many free products contradict the economist's assertion that "There is no free lunch"? No! Resources are used to produce each of these products, and because those resources have alternative uses, society gives up something else to get the "free" good. Because alternatives must be forsaken, there is no such thing as a free lunch.

So why are these goods offered for free? In a word: marketing! Firms sometimes offer free products to entice people to try them, hoping they will then purchase those goods later. Getting to try out the free version of an app may eventually entice you to buy the pay version that has more features. In other cases, a product is free only in conjunction with a larger purchase. To get the free bottle of soda, you must buy the large pizza. To get the free cell phone, you need to sign up for a year's worth of cell phone service.

But while "free" products may come at no cost to the individuals receiving them, they are never free to society because their manufacture requires the use of resources that could have been put to alternative uses.

ORIGIN OF THE IDEA

O1.2
Utility

satisfaction. Because they weigh costs and benefits, their economic decisions are "purposeful" or "rational," not "random" or "chaotic."

Consumers are purposeful in deciding what goods and services to buy. Business firms are purposeful in deciding what products to produce and how to produce them. Government entities are purposeful in deciding what public services to provide and how to finance them.

"Purposeful behavior" does not assume that people and institutions are immune from faulty logic and therefore are perfect decision makers. They sometimes make mistakes. Nor does it mean that people's decisions are unaffected by emotion or the decisions of those around them. Indeed, economists acknowledge that people are sometimes impulsive or emulative. "Purposeful behavior" simply means that people make decisions with some desired outcome in mind.

Rational self-interest is not the same as selfishness. In the economy, increasing one's own wage, rent, interest, or profit normally requires identifying and satisfying *somebody else's* wants! Also, people make personal sacrifices for others. They contribute time and money to charities because they derive pleasure from doing so. Parents help pay for their children's education for the same reason. These self-interested, but unselfish, acts help maximize the givers' satisfaction as much as any personal purchase of goods or services. Self-interested behavior is simply behavior designed to increase personal satisfaction, however it may be derived.

Marginal Analysis: Comparing Benefits and Costs

The economic perspective focuses largely on **marginal analysis**—comparisons of marginal benefits and marginal costs, usually for decision making. To economists, "marginal" means "extra," "additional," or "a change in." Most choices or decisions involve changes in the status quo, meaning the existing state of affairs.

Should you attend school for another year? Should you study an extra hour for an exam? Should you supersize your fries? Similarly, should a business expand or reduce its output? Should government increase or decrease its funding for a missile defense system?

ORIGIN OF THE IDEA

O1.3
Marginal analysis

Each option involves marginal benefits and, because of scarce resources, marginal costs. In making choices rationally, the decision maker must compare those two amounts. Example: You and your fiancée are shopping for an engagement ring. Should you buy a $\frac{1}{2}$-carat diamond, a $\frac{3}{4}$-carat diamond, a 1-carat diamond, or something even larger? The marginal cost of a larger-size diamond is the added expense beyond

CONSIDER THIS . . .

Fast-Food Lines

The economic perspective is useful in analyzing all sorts of behaviors. Consider an everyday example: the behavior of fast-food customers. When customers enter the restaurant, they go to the shortest line, believing that line will minimize their time cost of obtaining food. They are acting purposefully; time is limited, and people prefer using it in some way other than standing in line.

If one fast-food line is temporarily shorter than other lines, some people will move to that line. These movers apparently view the time saving from the shorter line (marginal benefit) as exceeding the cost of moving from their present line (marginal cost). The line switching tends to equalize line lengths. No further movement of customers between lines occurs once all lines are about equal.

Fast-food customers face another cost-benefit decision when a clerk opens a new station at the counter. Should they move to the new station or stay put? Those who shift to the new line decide that the time saving from the move exceeds the extra cost of physically moving. In so deciding, customers must also consider just how quickly they can get to the new station compared with others who may be contemplating the same move. (Those who hesitate in this situation are lost!)

Customers at the fast-food establishment do not have perfect information when they select lines. Thus, not all decisions turn out as expected. For example, you might enter a short line only to find that someone in front of you is ordering hamburgers and fries for 40 people in the Greyhound bus parked out back (and also that the guy taking orders in your new line is a trainee)! Nevertheless, at the time you made your decision, you thought it was optimal.

Finally, customers must decide what food to order when they arrive at the counter. In making their choices, they again compare marginal costs and marginal benefits in attempting to obtain the greatest personal satisfaction for their expenditure.

Economists believe that what is true for the behavior of customers at fast-food restaurants is true for economic behavior in general. Faced with an array of choices, consumers, workers, and businesses rationally compare marginal costs and marginal benefits when making decisions.

the cost of the smaller-size diamond. The marginal benefit is the perceived lifetime pleasure (utility) from the larger-size stone. If the marginal benefit of the larger diamond exceeds its marginal cost (and you can afford it), buy the larger stone. But if the marginal cost is more than the marginal benefit, you should buy the smaller diamond instead—even if you can afford the larger stone!

In a world of scarcity, the decision to obtain the marginal benefit associated with some specific option always includes the marginal cost of forgoing something else. The money spent on the larger-size diamond means forgoing some other product. An opportunity cost—the value of the next best thing forgone—is always present whenever a choice is made.

Theories, Principles, and Models

LO1.2 Describe the role of economic theory in economics.

Like the physical and life sciences, as well as other social sciences, economics relies on the **scientific method.** That procedure consists of several elements:

- Observing real-world behavior and outcomes.
- Based on those observations, formulating a possible explanation of cause and effect (hypothesis).
- Testing this explanation by comparing the outcomes of specific events to the outcome predicted by the hypothesis.
- Accepting, rejecting, and modifying the hypothesis, based on these comparisons.
- Continuing to test the hypothesis against the facts. If favorable results accumulate, the hypothesis evolves into a theory. A very well-tested and widely accepted theory is referred to as an economic law or an **economic principle**—a statement about economic behavior or the economy that enables prediction of the probable effects of certain actions. Combinations of such laws or principles are incorporated into models, which are simplified representations of how something works, such as a market or segment of the economy.

Economists develop theories of the behavior of individuals (consumers, workers) and institutions (businesses, governments) engaged in the production, exchange, and consumption of goods and services. Theories, principles, and models are "purposeful simplifications." The full scope of economic reality itself is too complex and bewildering to be understood as a whole. In developing theories, principles, and models economists remove the clutter and simplify.

Economic principles and models are highly useful in analyzing economic behavior and understanding how the economy operates. They are the tools for ascertaining cause and effect (or action and outcome) within the economic system. Good theories do a good job of explaining and predicting. They are supported by facts concerning how individuals and institutions actually behave in producing, exchanging, and consuming goods and services.

There are some other things you should know about economic principles.

- ***Generalizations*** Economic principles are generalizations relating to economic behavior or to the economy itself. Economic principles are expressed as the tendencies of typical or average consumers, workers, or business firms. For example, economists say that consumers buy more of a particular product when its price falls. Economists recognize that some consumers may increase their purchases by a large amount, others by a small amount, and a few not at all. This "price-quantity" principle, however, holds for the typical consumer and for consumers as a group.
- ***Other-things-equal assumption*** In constructing their theories, economists use the *ceteris paribus* or **other-things-equal assumption**—the assumption that factors other than those being considered do not change. They assume that all variables except those under immediate consideration are held constant for a particular analysis. For example, consider the relationship between the price of Pepsi and the amount of it purchased. Assume that of all the factors that might influence the amount of Pepsi purchased (for example, the price of Pepsi, the price of Coca-Cola, and consumer incomes and preferences), only the price of Pepsi varies. This is helpful because the economist can then focus on the relationship between the price of Pepsi and purchases of Pepsi in isolation without being confused by changes in other variables.

ORIGIN OF THE IDEA

O1.4

Ceteris paribus

- ***Graphical expression*** Many economic models are expressed graphically. Be sure to read the special appendix at the end of this chapter as a review of graphs.

Microeconomics and Macroeconomics

LO1.3 Distinguish microeconomics from macroeconomics and positive economics from normative economics.

Economists develop economic principles and models at two levels.

Microeconomics

Microeconomics is the part of economics concerned with decision making by individual customers, workers, households, and business firms. At this level of analysis, we observe the details of their behavior under a figurative microscope. We measure the price of a specific product, the number of workers employed by a single firm, the revenue or income of a particular firm or household, or the expenditures of a specific firm, government entity, or family. In microeconomics, we examine the sand, rocks, and shells, not the beach.

Macroeconomics

Macroeconomics examines the performance and behavior of the economy as a whole. It focuses its attention on economic growth, the business cycle, interest rates, inflation, and the behavior of major economic aggregates such as the government, household, and business sectors. An **aggregate** is a collection of specific economic units treated as if they were one unit. Therefore, we might lump together the millions of consumers in the U.S. economy and treat them as if they were one huge unit called "consumers."

In using aggregates, macroeconomics seeks to obtain an overview, or general outline, of the structure of the economy and the relationships of its major aggregates. Macroeconomics speaks of such economic measures as total output, total employment, total income, aggregate expenditures, and the general level of prices in analyzing various economic problems. Very little attention is given to the specific units making up the various aggregates.

Figuratively, macroeconomics looks at the beach, not the pieces of sand, the rocks, and the shells.

The micro–macro distinction does not mean that economics is so highly compartmentalized that every topic can be readily labeled as either micro or macro; many topics and subdivisions of economics are rooted in both. Example: While the problem of unemployment is usually treated as a macroeconomic topic (because unemployment relates to aggregate production), economists recognize that the decisions made by *individual* workers on how long to search for jobs and the way *specific* labor markets encourage or impede hiring are also critical in determining the unemployment rate.

Positive and Normative Economics

Both microeconomics and macroeconomics contain elements of positive economics and normative economics. **Positive economics** focuses on facts and cause-and-effect relationships. It includes description, theory development, and theory testing. Positive economics avoids value judgments. It tries to establish scientific statements about economic behavior and deals with what the economy is actually like. Such scientific-based analysis is critical to good policy analysis.

Economic policy, on the other hand, involves **normative economics,** which incorporates value judgments about what the economy should be like or what particular policy actions should be recommended to achieve a desirable goal. Normative economics looks at the desirability of certain aspects of the economy. It underlies expressions of support for particular economic policies.

Positive economics concerns *what is*, whereas normative economics embodies subjective feelings about *what ought to be*. Examples: Positive statement: "The unemployment rate in France is higher than that in the United States." Normative statement: "France ought to undertake policies to make its labor market more flexible to reduce unemployment rates." Whenever words such as "ought" or "should" appear in a sentence, you are very likely encountering a normative statement.

Most of the disagreement among economists involves normative, value-based policy questions. Of course, economists sometime disagree about which theories or models best represent the economy and its parts, but they agree on a full range of economic principles. Most economic controversy thus reflects differing opinions or value judgments about what society should be like.

QUICK REVIEW 1.1

- Economics examines how individuals, institutions, and society make choices under conditions of scarcity.
- The economic perspective stresses (a) resource scarcity and the necessity of making choices, (b) the assumption of purposeful (or rational) behavior, and (c) comparisons of marginal benefit and marginal cost.
- In choosing the best option, people incur an opportunity cost—the value of the next-best option.
- Economists use the scientific method to establish economic theories—cause-effect generalizations about the economic behavior of individuals and institutions.
- Microeconomics focuses on specific decision-making units within the economy. Macroeconomics examines the economy as a whole.
- Positive economics deals with factual statements ("what is"); normative economics involves value judgments ("what ought to be").

Individual's Economizing Problem

LO1.4 Explain the individual's economizing problem and how trade-offs, opportunity costs, and attainable combinations can be illustrated with budget lines.

A close examination of the **economizing problem**—the need to make choices because economic wants exceed economic means—will enhance your understanding of economic models and the difference between microeconomic and macroeconomic analysis. Let's first build a microeconomic model of the economizing problem faced by an individual.

Limited Income

We all have a finite amount of income, even the wealthiest among us. Even Donald Trump must decide how to spend his money! And the majority of us have much more limited means. Our income comes to us in the form of wages, interest, rent, and profit, although we may also receive money from government programs or family members. As Global Perspective 1.1 shows, the average income of Americans in 2011 was $48,450. In the poorest nations, it was less than $500.

Unlimited Wants

For better or worse, most people have virtually unlimited wants. We desire various goods and services that provide utility. Our wants extend over a wide range of products, from *necessities* (for example, food, shelter, and clothing) to *luxuries* (for example, perfumes, yachts, and sports cars). Some wants such as basic food, clothing, and shelter have biological roots. Other wants, for example, specific kinds of food, clothing, and shelter, arise from the conventions and customs of society.

Over time, as new and improved products are introduced, economic wants tend to change and multiply. Only recently have people wanted iPods, Internet service, or camera phones because those products did not exist a few decades ago. Also, the satisfaction of certain wants may trigger others: the acquisition of a Ford Focus or a Honda Civic has been known to whet the appetite for a Lexus or a Mercedes.

Services, as well as goods, satisfy our wants. Car repair work, the removal of an inflamed appendix, legal and accounting advice, and haircuts all satisfy human wants. Actually, we buy many goods, such as automobiles and washing machines, for the services they render. The differences between goods and services are often smaller than they appear to be.

For most people, the desires for goods and services cannot be fully satisfied. Bill Gates may have all that he wants for himself, but his massive charitable giving suggests that he keenly wants better health care for the world's poor. Our desires for a particular good or service can be satisfied; over a short period of time we can surely get enough toothpaste or pasta. And one appendectomy is plenty. But our broader desire for more goods and services and higher-quality goods and services seems to be another story.

Because we have only limited income (usually through our work) but seemingly insatiable wants, it is in our self-interest to economize: to pick and choose goods and services that maximize our satisfaction given the limitations we face.

GLOBAL PERSPECTIVE 1.1

Average Income, Selected Nations

Average income (total income/population) and therefore typical individual budget constraints vary greatly among nations.

Country	Per Capita Income, 2011 (U.S. dollars, based on exchange rates)
Norway	$88,890
Switzerland	76,380
United States	48,450
Singapore	42,930
France	42,420
South Korea	20,870
Mexico	9240
China	4940
Iraq	2640
India	1410
Madagascar	430
Congo	190

Source: World Bank, **www.worldbank.org**.

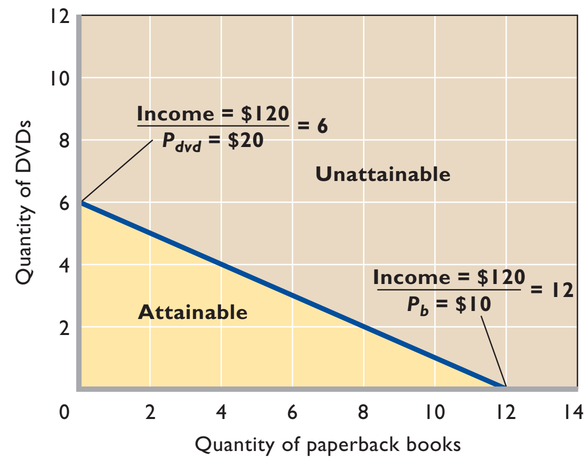
A Budget Line

We can clarify the economizing problem facing consumers by visualizing a **budget line** (or, more technically, a *budget constraint*). It is a schedule or curve that shows various combinations of two products a consumer can purchase with a specific money income. Although we assume two products, the analysis generalizes to the full range of products available to consumers.

FIGURE 1.1 A consumer's budget line. The budget line (or budget constraint) shows all the combinations of any two products that can be purchased, given the prices of the products and the consumer's money income.

The Budget Line: Whole-Unit Combinations of DVDs and Paperback Books Attainable with an Income of $120

Units of DVDs (Price = $20)	Units of Books (Price = $10)	Total Expenditure
6	0	$120 (= $120 + $0)
5	2	$120 (= $100 + $20)
4	4	$120 (= $80 + $40)
3	6	$120 (= $60 + $60)
2	8	$120 (= $40 + $80)
1	10	$120 (= $20 + $100)
0	12	$120 (= $0 + $120)



To understand the idea of a budget line, suppose that you received a Barnes & Noble gift card as a birthday present. The $120 card is soon to expire. You take the card to the store and confine your purchase decisions to two alternatives: DVDs and paperback books. DVDs are $20 each and paperback books are $10 each. Your purchase options are shown in the table in Figure 1.1.

At one extreme, you might spend all of your $120 "income" on 6 DVDs at $20 each and have nothing left to spend on books. Or, by giving up 2 DVDs and thereby gaining $40, you can have 4 DVDs at $20 each and 4 books at $10 each. And so on to the other extreme, at which you could buy 12 books at $10 each, spending your entire gift card on books with nothing left to spend on DVDs.

The graph in Figure 1.1 shows the budget line. Note that the graph is not restricted to whole units of DVDs and books as is the table. Every point on the graph represents a possible combination of DVDs and books, including fractional quantities. The slope of the graphed budget line measures the ratio of the price of books (P_b) to the price of DVDs (P_{dvd}); more precisely, the slope is $P_b/P_{dvd} = \$-10/\$+20 = -\frac{1}{2}$. So you must forgo 1 DVD (measured on the vertical axis) to buy 2 books (measured on the horizontal axis). This yields a slope of $-\frac{1}{2}$ or $-.5$.

The budget line illustrates several ideas.

Attainable and Unattainable Combinations All the combinations of DVDs and books on or inside the budget line are *attainable* from the $120 of money income. You can afford to buy, for example, 3 DVDs at $20 each and 6 books at $10 each. You also can obviously afford to buy 2 DVDs and 5 books, thereby using up only $90 of the $120 available on your gift card. But to achieve maximum utility you will want to spend the full $120. The budget line shows all combinations that cost exactly the full $120.

In contrast, all combinations beyond the budget line are *unattainable*. The $120 limit simply does not allow you to purchase, for example, 5 DVDs at $20 each and 5 books at $10 each. That $150 expenditure would clearly exceed the $120 limit. In Figure 1.1 the attainable combinations are on and within the budget line; the unattainable combinations are beyond the budget line.

Trade-Offs and Opportunity Costs The budget line in Figure 1.1 illustrates the idea of trade-offs arising from limited income. To obtain more DVDs, you have to give up some books. For example, to obtain the first DVD, you trade off 2 books. So the opportunity cost of the first DVD is 2 books. To obtain the second DVD the opportunity cost is also 2 books. The straight-line budget constraint, with its constant slope, indicates constant opportunity cost. That is, the opportunity cost of 1 extra DVD remains the same (= 2 books) as more DVDs are purchased. And, in reverse, the opportunity cost of 1 extra book does not change (= $\frac{1}{2}$ DVD) as more books are bought.

ORIGIN OF THE IDEA

O1.5

Opportunity costs

Choice Limited income forces people to choose what to buy and what to forgo to fulfill wants. You will select the combination of DVDs and paperback books that you think is "best." That is, you will evaluate your marginal benefits and marginal costs (here, product price) to make choices that maximize your satisfaction. Other people, with the same $120 gift card, would undoubtedly make different choices.

Income Changes The location of the budget line varies with money income. An increase in money income shifts the budget line to the right; a decrease in money income shifts it to the left. To verify this, recalculate the table in Figure 1.1, assuming the card value (income) is (a) $240 and (b) $60, and plot the new budget lines in the graph. No wonder people like to have more income: That shifts their budget lines outward and enables them to buy more goods and services. But even with more income, people will still face spending trade-offs, choices, and opportunity costs.

WORKED PROBLEMS

W1.1

Budget lines

CONSIDER THIS . . .

Did Zuckerberg, Winfrey, and James Make Bad Choices?

Opportunity costs come into play in decisions well beyond simple buying decisions. Consider the different choices people make with respect to college. The average salaries earned by college graduates are nearly twice as high as those earned by persons with just high school diplomas. For most capable students, "Go to college, stay in college, and earn a degree" is very sound advice.

Yet Facebook founder Mark Zuckerberg and talk show host Oprah Winfrey* both dropped out of college, and basketball star LeBron James never even bothered to start classes. What were they thinking? Unlike most students, Zukerberg faced enormous opportunity costs for staying in college. He had a vision for his company, and dropping out helped to ensure Facebook's success. Similarly, Winfrey landed a spot in local television news when she was a teenager, eventually producing and starring in the Oprah Winfrey Show when she was 32 years old. Getting a degree in her twenties might have interrupted the string of successes that made her famous talk show possible. And James knew that professional athletes have short careers. Therefore, going to college directly after high school would have taken away four years of his peak earning potential.

So Zuckerberg, Winfrey, and James understood opportunity costs and made their choices accordingly. The size of opportunity costs matters greatly in making individual decisions.

*Winfrey eventually went back to school and earned a degree from Tennessee State University when she was in her thirties.

QUICK REVIEW 1.2

- Because wants exceed incomes, individuals face an economizing problem; they must decide what to buy and what to forgo.
- A budget line (budget constraint) shows the various combinations of two goods that a consumer can purchase with a specific money income.
- Straight-line budget constraints imply constant opportunity costs for both goods.

Society's Economizing Problem

LO1.5 List the categories of scarce resources and delineate the nature of society's economizing problem.

Society must also make choices under conditions of scarcity. It, too, faces an economizing problem. Should it devote more of its limited resources to the criminal justice system (police, courts, and prisons) or to education (teachers, books, and schools)? If it decides to devote more resources to both, what other goods and services does it forgo? Health care? Energy development?

Scarce Resources

Society has limited or scarce **economic resources,** meaning all natural, human, and manufactured resources that go into the production of goods and services. This includes the entire set of factory and farm buildings and all the equipment, tools, and machinery used to produce manufactured goods and agricultural products; all transportation and communication facilities; all types of labor; and land and mineral resources.

Resource Categories

Economists classify economic resources into four general categories.

Land Land means much more to the economist than it does to most people. To the economist **land** includes all natural resources ("gifts of nature") used in the production process. These include forests, mineral and oil deposits, water resources, wind power, sunlight, and arable land.

Labor The resource **labor** consists of the physical actions and mental activities that people contribute to the production of goods and services. The work-related activities of a logger, retail clerk, machinist, teacher, professional football player, and nuclear physicist all fall under the general heading "labor."

Capital For economists, **capital** (or capital goods) includes all manufactured aids used in producing consumer goods and services. Included are all factory, storage, transportation, and distribution facilities, as well as tools and machinery. Economists use the term **investment** to describe spending that pays for the production and accumulation of capital goods.

Capital goods differ from consumer goods because consumer goods satisfy wants directly, whereas capital goods do so indirectly by aiding the production of consumer goods. For example, large commercial baking ovens (capital goods) help make loaves of bread (consumer goods). Note that the term "capital" as used by economists refers not to money but to tools, machinery, and other productive equipment. Because money produces nothing, economists do not include it as an economic resource. Money (or money capital or financial capital) is simply a means for purchasing goods and services, including capital goods.

Entrepreneurial Ability Finally, there is the special human resource, distinct from labor, called **entrepreneurial ability.** It is supplied by **entrepreneurs,** who perform several critically important economic functions:

- The entrepreneur takes the initiative in combining the resources of land, labor, and capital to produce a good or a service. Both a sparkplug and a catalyst, the entrepreneur is the driving force behind production and the agent who combines the other resources in what is hoped will be a successful business venture.
- The entrepreneur makes the strategic business decisions that set the course of an enterprise.
- The entrepreneur innovates. He or she commercializes new products, new production techniques, or even new forms of business organization.
- The entrepreneur bears risk. Innovation is risky, as nearly all new products and ideas are subject to the possibility of failure as well as success. Progress would cease without entrepreneurs who are willing to take on risk by devoting their time, effort, and ability—as well as their own money and the money of others—to commercializing new products and ideas that may enhance society's standard of living.

Because land, labor, capital, and entrepreneurial ability are combined to produce goods and services, they are called the **factors of production,** or simply "inputs."

Production Possibilities Model

LO1.6 Apply production possibilities analysis, increasing opportunity costs, and economic growth.

Society uses its scarce resources to produce goods and services. The alternatives and choices it faces can best be understood through a macroeconomic model of production possibilities. To keep things simple, let's initially assume

- ***Full employment*** The economy is employing all of its available resources.
- ***Fixed resources*** The quantity and quality of the factors of production are fixed.
- ***Fixed technology*** The state of technology (the methods used to produce output) is constant.
- ***Two goods*** The economy is producing only two goods: pizzas and industrial robots. Pizzas symbolize **consumer goods,** products that satisfy our wants directly; industrial robots (for example, the kind used to weld automobile frames) symbolize **capital goods,** products that satisfy our wants indirectly by making possible more efficient production of consumer goods.

Production Possibilities Table

A production possibilities table lists the different combinations of two products that can be produced with a specific set of resources, assuming full employment. Table 1.1 presents a simple, hypothetical economy that is producing pizzas and industrial robots; the data are, of course, hypothetical. At alternative A, this economy would be devoting all its available resources to the production of industrial robots (capital goods); at alternative E, all resources would go to pizza production (consumer goods). Those alternatives are unrealistic extremes; an economy typically produces both capital goods and consumer goods, as in B, C, and D. As we move from alternative A to E, we increase the production of pizzas at the expense of the production of industrial robots.

Because consumer goods satisfy our wants directly, any movement toward E looks tempting. In producing more pizzas, society increases the satisfaction of its current wants. But there is a cost: More pizzas mean fewer industrial robots. This shift of resources to consumer goods catches up with society over time because the stock of capital goods expands more slowly, thereby reducing potential future production. By moving toward alternative E, society chooses "more now" at the expense of "much more later."

By moving toward A, society chooses to forgo current consumption, thereby freeing up resources that can be used to increase the production of capital goods. By building up

TABLE 1.1 **Production Possibilities of Pizzas and Industrial Robots**

	Production Alternatives				
Type of Product	A	B	C	D	E
Pizzas (in hundred thousands)	0	1	2	3	4
Robots (in thousands)	10	9	7	4	0

KEY GRAPH

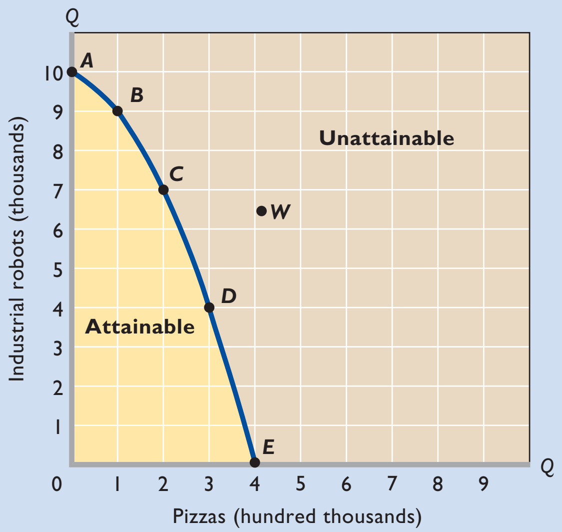


FIGURE 1.2 The production possibilities curve. Each point on the production possibilities curve represents some maximum combination of two products that can be produced if resources are fully employed. When an economy is operating on the curve, more industrial robots means fewer pizzas, and vice versa. Limited resources and a fixed technology make any combination of industrial robots and pizzas lying outside the curve (such as at *W*) unattainable. Points inside the curve are attainable, but they indicate that full employment is not being realized.

QUICK QUIZ FOR FIGURE 1.2

1. Production possibilities curve *ABCDE* is bowed out from the origin because:
- **a.** the marginal benefit of pizzas declines as more pizzas are consumed.
- **b.** the curve gets steeper as we move from *E* to *A*.
- **c.** it reflects the law of increasing opportunity costs.
- **d.** resources are scarce.

2. The marginal opportunity cost of the second unit of pizza is:
- **a.** 2 units of robots.
- **b.** 3 units of robots.
- **c.** 7 units of robots.
- **d.** 9 units of robots.

3. The total opportunity cost of 7 units of robots is:
- **a.** 1 unit of pizza.
- **b.** 2 units of pizza.
- **c.** 3 units of pizza.
- **d.** 4 units of pizza.

4. All points on this production possibilities curve necessarily represent:
- **a.** society's optimal choice.
- **b.** less than full use of resources.
- **c.** unattainable levels of output.
- **d.** full employment.

Answers: 1. c; 2. a; 3. b; 4. d

its stock of capital this way, society will have greater future production and, therefore, greater future consumption. By moving toward A, society is choosing "more later" at the cost of "less now."

Generalization: At any point in time, a fully employed economy must sacrifice some of one good to obtain more of another good. Scarce resources prohibit a fully employed economy from having more of both goods. Society must choose among alternatives. There is no such thing as a free pizza, or a free industrial robot. Having more of one thing means having less of something else.

Production Possibilities Curve

The data presented in a production possibilities table are shown graphically as a **production possibilities curve.** Such a curve displays the different combinations of goods and services that society can produce in a fully employed economy, assuming a fixed availability of supplies of resources and fixed technology. We arbitrarily represent the economy's output of capital goods (here, industrial robots) on the vertical axis and the output of consumer goods (here, pizzas) on the horizontal axis, as shown in **Figure 1.2 (Key Graph).**

Each point on the production possibilities curve represents some maximum output of the two products. The curve is a "constraint" because it shows the limit of attainable outputs. Points on the curve are attainable as long as the economy uses all its available resources. Points lying inside the curve are also attainable, but they reflect less total output and therefore are not as desirable as points on the curve. Points inside the curve imply that the economy could have more of both industrial robots and pizzas if it achieved full employment of its resources. Points lying beyond the production possibilities curve, like *W*, would represent a greater output than the output at any point on the curve. Such points, however, are unattainable with the current availability of resources and technology.

Law of Increasing Opportunity Costs

Figure 1.2 clearly shows that more pizzas mean fewer industrial robots. The number of units of industrial robots that must be given up to obtain another unit of pizzas, of course, is the opportunity cost of that unit of pizzas.

In moving from alternative A to alternative B in Table 1.1, the cost of 1 additional unit of pizzas is 1 fewer unit of industrial robots. But when additional units are considered—B to C, C to D, and D to E—an important economic principle is revealed: For society, the opportunity cost of each additional unit of pizzas is greater than the opportunity cost of the preceding one. When we move from A to B, just 1 unit of industrial robots is sacrificed for 1 more unit of pizzas; but in going from B to C we sacrifice 2 additional units of industrial robots for 1 more unit of pizzas; then 3 more of industrial robots for 1 more of pizzas; and finally 4 for 1. Conversely, confirm that as we move from E to A, the cost of an additional unit of industrial robots (on average) is $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, and 1 unit of pizzas, respectively, for the four successive moves.

Our example illustrates the **law of increasing opportunity costs.** As the production of a particular good increases, the opportunity cost of producing an additional unit rises.

Shape of the Curve The law of increasing opportunity costs is reflected in the shape of the production possibilities curve: The curve is bowed out from the origin of the graph. Figure 1.2 shows that when the economy moves from *A* to *E*, it must give up successively larger amounts of industrial robots (1, 2, 3, and 4) to acquire equal increments of pizzas (1, 1, 1, and 1). This is shown in the slope of the production possibilities curve, which becomes steeper as we move from *A* to *E*.

Economic Rationale The law of increasing opportunity costs is driven by the fact that economic resources are not completely adaptable to alternative uses. Many resources are better at producing one type of good than at producing others. Consider land. Some land is highly suited to growing the ingredients necessary for pizza production. But as pizza production expands, society has to start using land that is less bountiful for farming. Other land is rich in mineral deposits and therefore well-suited to producing the materials needed to make industrial robots. That land will be the first land devoted to the production of industrial robots. But as society steps up the production of robots, it must use land that is less and less suited to making their components.

If we start at *A* and move to *B* in Figure 1.2, we can shift resources whose productivity is relatively high in pizza production and low in industrial robots. But as we move from *B* to *C*, *C* to *D*, and so on, resources highly productive in pizzas become increasingly scarce. To get more pizzas, resources whose productivity in industrial robots is relatively great will be needed. Increasingly more of such resources, and hence greater sacrifices of industrial robots, will be needed to achieve each 1-unit increase in pizzas. This lack of perfect flexibility, or interchangeability, on the part of resources is the cause of increasing opportunity costs for society.

WORKED PROBLEMS

W1.2
Production possibilities

Optimal Allocation

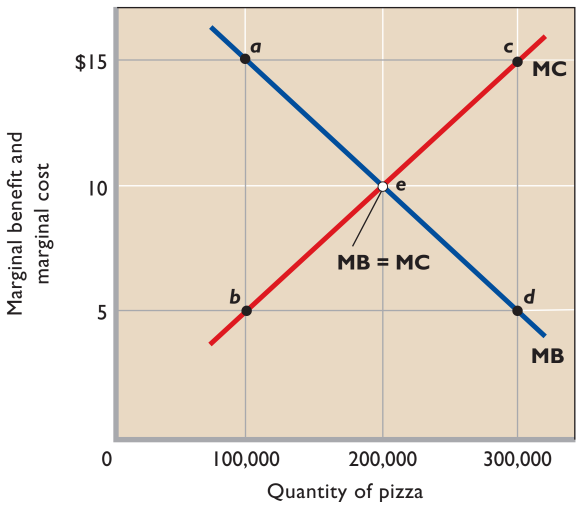
Of all the attainable combinations of pizzas and industrial robots on the curve in Figure 1.2, which is optimal (best)? That is, what specific quantities of resources should be allocated to pizzas and what specific quantities should be allocated to industrial robots in order to maximize satisfaction?

Recall that economic decisions center on comparisons of marginal benefit (MB) and marginal cost (MC). Any economic activity should be expanded as long as marginal benefit exceeds marginal cost and should be reduced if marginal cost exceeds marginal benefit. The optimal amount of the activity occurs where MB = MC. Society needs to make a similar assessment about its production decision.

Consider pizzas. We already know from the law of increasing opportunity costs that the marginal cost of additional units of pizza will rise as more units are produced. At the same time, we need to recognize that the extra or marginal benefits that come from producing and consuming pizza decline with each successive unit of pizza. Consequently, each successive unit of pizza brings with it both increasing marginal costs and decreasing marginal benefits.

The optimal quantity of pizza production is indicated by point *e* at the intersection of the MB and MC curves: 200,000 units in Figure 1.3. Why is this amount the optimal quantity? If only 100,000 units of pizzas were produced, the marginal benefit of an extra unit of pizza (point *a*) would exceed its marginal cost (point *b*). In money terms, MB is $15, while MC is only $5. When society gains something worth $15 at a marginal cost of only $5, it is better off. In Figure 1.3, net gains can continue to be realized until pizza-product production has been increased to 200,000.

FIGURE 1.3 Optimal output: MB = MC. Achieving the optimal output requires the expansion of a good's output until its marginal benefit (MB) and marginal cost (MC) are equal. No resources beyond that point should be allocated to the product. Here, optimal output occurs at point *e*, where 200,000 units of pizzas are produced.



In contrast, the production of 300,000 units of pizzas is excessive. There the MC of an added unit is $15 (point *c*) and its MB is only $5 (point *d*). This means that 1 unit of pizza is worth only $5 to society but costs it $15 to obtain. This is a losing proposition for society!

So resources are being efficiently allocated to any product when the marginal benefit and marginal cost of its output are equal (MB = MC). Suppose that by applying the same analysis to industrial robots, we find that the optimal (MB = MC) quantity of robots is 7,000. This would mean that alternative *C* (200,000 units of pizzas and 7,000 units of industrial robots) on the production possibilities curve in Figure 1.2 would be optimal for this economy.

QUICK REVIEW 1.3

- Economists categorize economic resources as land, labor, capital, and entrepreneurial ability.
- The production possibilities curve illustrates several ideas: (a) scarcity of resources is implied by the area of unattainable combinations of output lying outside the production possibilities curve; (b) choice among outputs is reflected in the variety of attainable combinations of goods lying along the curve; (c) opportunity cost is illustrated by the downward slope of the curve; (d) the law of increasing opportunity costs is reflected in the bowed-outward shape of the curve.
- A comparison of marginal benefits and marginal costs is needed to determine the best or optimal output mix on a production possibilities curve.

CONSIDER THIS . . .

The Economics of War

Production possibilities analysis is helpful in assessing the costs and benefits of waging the broad war on terrorism, including the wars in Afghanistan and Iraq. At the end of 2011, the estimated cost of these efforts exceeded $1.4 trillion.

If we categorize all U.S. production as either "defense goods" or "civilian goods," we can measure them on the axes of a production possibilities diagram such as that shown in Figure 1.2. The opportunity cost of using more resources for defense goods is the civilian goods sacrificed. In a fully employed economy, more defense goods are achieved at the opportunity cost of fewer civilian goods—health care, education, pollution control, personal computers, houses, and so on. The cost of war and defense is the other goods forgone. The benefits of these activities are numerous and diverse but clearly include the gains from protecting against future loss of American lives, assets, income, and well-being.

Society must assess the marginal benefit (MB) and marginal cost (MC) of additional defense goods to determine their optimal amounts—where to locate on the defense goods–civilian goods production possibilities curve. Although estimating marginal benefits and marginal costs is an imprecise art, the MB-MC framework is a useful way of approaching choices. An optimal allocation of resources requires that society expand production of defense goods until MB = MC.

The events of September 11, 2001, and the future threats they foreshadowed increased the marginal benefits of defense goods, as perceived by Americans. If we label the horizontal axis in Figure 1.3 "defense goods" and draw in a rightward shift of the MB curve, you will see that the optimal quantity of defense goods rises. In view of the concerns relating to September 11, the United States allocated more of its resources to defense. But the MB-MC analysis also reminds us we can spend too much on defense, as well as too little. The United States should not expand defense goods beyond the point where MB = MC. If it does, it will be sacrificing civilian goods of greater value than the defense goods obtained.

Unemployment, Growth, and the Future

LO1.7 Explain how economic growth and international trade increase consumption possibilities.

In the depths of the Great Depression of the 1930s, one-quarter of U.S. workers were unemployed and one-third of

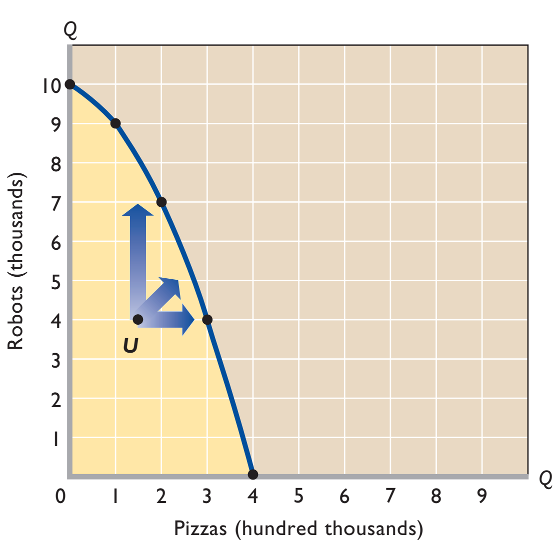
U.S. production capacity was idle. Subsequent downturns have been much less severe. During the deep 2007–2009 recession, for instance, production fell by a comparably smaller 3.7 percent and 1-in-10 workers was without a job.

Almost all nations have experienced widespread unemployment and unused production capacity from business downturns at one time or another. Since 2000, for example, many nations—including Argentina, Japan, Mexico, Germany, and South Korea—have had economic downturns and unemployment.

How do these realities relate to the production possibilities model? Our analysis and conclusions change if we relax the assumption that all available resources are fully employed. The five alternatives in Table 1.1 represent maximum outputs; they illustrate the combinations of pizzas and industrial robots that can be produced when the economy is operating at full employment. With unemployment, this economy would produce less than each alternative shown in the table.

Graphically, we represent situations of unemployment by points inside the original production possibilities curve (reproduced here in Figure 1.4). Point *U* is one such point. Here the economy is falling short of the various maximum combinations of pizzas and industrial robots represented by the points on the production possibilities curve. The arrows in Figure 1.4 indicate three possible paths back to full employment. A move toward full employment would yield a greater output of one or both products.

FIGURE 1.4 Unemployment and the production possibilities curve. Any point inside the production possibilities curve, such as *U*, represents unemployment or a failure to achieve full employment. The arrows indicate that by realizing full employment, the economy could operate on the curve. This means it could produce more of one or both products than it is producing at point *U*.



A Growing Economy

When we drop the assumptions that the quantity and quality of resources and technology are fixed, the production possibilities curve shifts positions and the potential maximum output of the economy changes.

Increases in Resource Supplies Although resource supplies are fixed at any specific moment, they change over time. For example, a nation's growing population brings about increases in the supplies of labor and entrepreneurial ability. Also, labor quality usually improves over time via more education and training. Historically, the economy's stock of capital has increased at a significant, though unsteady, rate. And although some of our energy and mineral resources are being depleted, new sources are also being discovered. The development of irrigation systems, for example, adds to the supply of arable land.

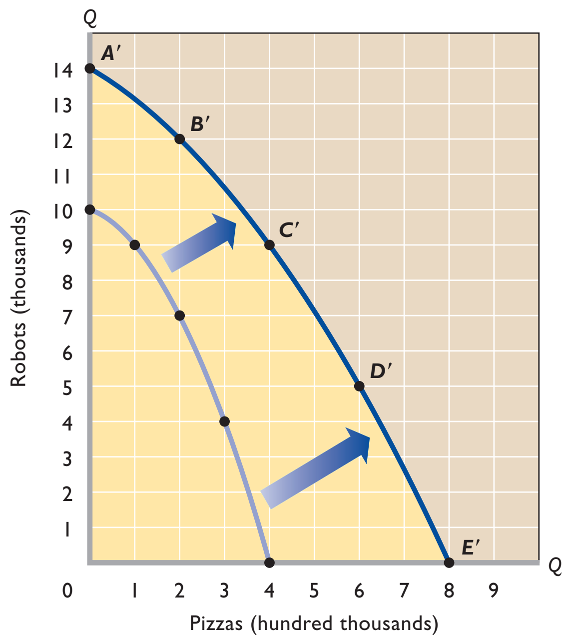
The net result of these increased supplies of the factors of production is the ability to produce more of both consumer goods and capital goods. Thus, 20 years from now, the production possibilities may supersede those shown in Table 1.1. The new production possibilities might look like those in the table in Figure 1.5. The greater abundance of resources will result in a greater potential output of one or both products at each alternative. The economy will have achieved economic growth in the form of expanded potential output. Thus, when an increase in the quantity or quality of resources occurs, the production possibilities curve shifts outward and to the right, as illustrated by the move from the inner curve to curve *A′B′C′D′E′* in Figure 1.5. This sort of shift represents growth of economic capacity, which, when used, means **economic growth:** a larger total output.

Advances in Technology An advancing technology brings both new and better goods and improved ways of producing them. For now, let's think of technological advance as being only improvements in the methods of production, for example, the introduction of computerized systems to manage inventories and schedule production. These advances alter our previous discussion of the economizing problem by allowing society to produce more goods with available resources. As with increases in resource supplies, technological advances make possible the production of more industrial robots *and* more pizzas.

A real-world example of improved technology is the recent surge of new technologies relating to computers, communications, and biotechnology. Technological advances have dropped the prices of computers and greatly

FIGURE 1.5 Economic growth and the production possibilities curve. The increase in supplies of resources, improvements in resource quality, and technological advances that occur in a dynamic economy move the production possibilities curve outward and to the right, allowing the economy to have larger quantities of both types of goods.

	Production Alternatives				
Type of Product	**A′**	**B′**	**C′**	**D′**	**E′**
Pizzas (in hundred thousands)	0	2	4	6	8
Robots (in thousands)	14	12	9	5	0



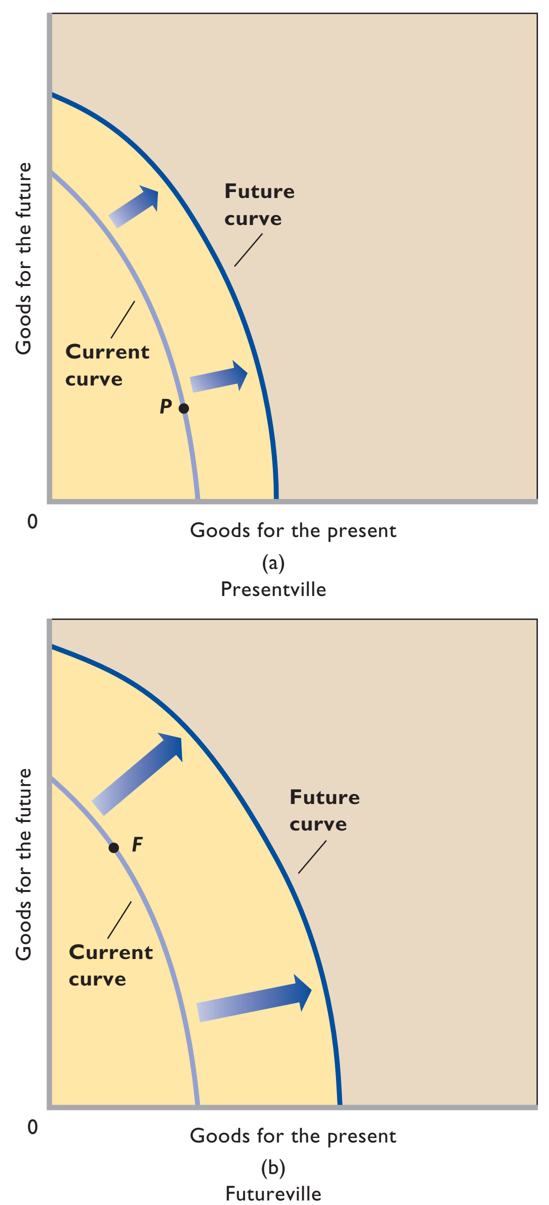
increased their speed. Improved software has greatly increased the everyday usefulness of computers. Cellular phones and the Internet have increased communications capacity, enhancing production and improving the efficiency of markets. Advances in biotechnology have resulted in important agricultural and medical discoveries. These and other new and improved technologies have contributed to U.S. economic growth (outward shifts of the nation's production possibilities curve).

Conclusion: Economic growth is the result of (1) increases in supplies of resources, (2) improvements in resource quality, and (3) technological advances. The consequence of growth is that a full-employment economy can enjoy a greater output of both consumption goods and capital goods. Whereas static, no-growth economies must sacrifice some of one good to obtain more of another, dynamic, growing economies can have larger quantities of both goods.

Present Choices and Future Possibilities

An economy's current choice of positions on its production possibilities curve helps determine the future location of that curve. Let's designate the two axes of the production possibilities curve as "goods for the future" and "goods for the present," as in Figure 1.6. Goods for the future are such

FIGURE 1.6 Present choices and future locations of production possibilities curves. (a) Presentville's current choice to produce more "present goods" and fewer "future goods," as represented by point *P*, will result in a modest outward shift of the production possibilities curve in the future. (b) Futureville's current choice of producing fewer "present goods" and more "future goods," as depicted by point *F*, will lead to a greater outward shift of the production possibilities curve in the future.



LAST WORD

Pitfalls to Sound Economic Reasoning

Because They Affect Us So Personally, We Often Have Difficulty Thinking Accurately and Objectively About Economic Issues.

Here are some common pitfalls to avoid in successfully applying the economic perspective.

Biases Most people bring a bundle of biases and preconceptions to the field of economics. For example, some might think that corporate profits are excessive or that lending money is always superior to borrowing money. Others might believe that government is necessarily less efficient than businesses or that more government regulation is always better than less. Biases cloud thinking and interfere with objective analysis. All of us must be willing to shed biases and preconceptions that are not supported by facts.

Loaded Terminology The economic terminology used in newspapers and broadcast media is sometimes emotionally biased, or loaded. The writer or spokesperson may have a cause to promote or an ax to grind and may slant comments accordingly. High profits may be labeled "obscene," low wages may be called "exploitative," or self-interested behavior may be "greed." Government workers may be referred to as "mindless bureaucrats" and those favoring stronger government regulations may be called "socialists." To objectively analyze economic issues, you must be prepared to reject or discount such terminology.

Fallacy of Composition Another pitfall in economic thinking is the assumption that what is true for one individual or part of a whole is necessarily true for a group of individuals or the whole. This is a logical fallacy called the *fallacy of composition;* the assumption is not correct. A statement that is valid for an individual or part is not necessarily valid for the larger group or whole. Noneconomic example: You may see the action better if

things as capital goods, research and education, and preventive medicine. They increase the quantity and quality of property resources, enlarge the stock of technological information, and improve the quality of human resources. As we have already seen, goods for the future such as capital goods are the ingredients of economic growth. Goods for the present are consumer goods such as food, clothing, and entertainment.

Now suppose there are two hypothetical economies, Presentville and Futureville, that are initially identical in every respect except one: Presentville's current choice of positions on its production possibilities curve strongly favors present goods over future goods. Point P in Figure 1.6a indicates that choice. It is located quite far down the curve to the right, indicating a high priority for goods for the present, at the expense of less goods for the future. Futureville, in contrast, makes a current choice that stresses larger amounts of future goods and smaller amounts of present goods, as shown by point F in Figure 1.6b.

Now, other things equal, we can expect Futureville's future production possibilities curve to be farther to the right than Presentville's future production possibilities curve. By currently choosing an output more favorable to technological advances and to increases in the quantity and quality of resources, Futureville will achieve greater economic growth than Presentville. In terms of capital goods, Futureville is choosing to make larger current additions to its "national factory" by devoting more of its current output to capital than does Presentville. The you leap to your feet to see an outstanding play at a football game. But if all the spectators leap to their feet at the same time, nobody—including you—will have a better view than when all remained seated.

Here are two economic examples: An individual stockholder can sell shares of, say, Google stock without affecting the price of the stock. The individual's sale will not noticeably reduce the share price because the sale is a negligible fraction of the total shares of Google being bought and sold. But if all the Google shareholders decide to sell their shares the same day, the market will be flooded with shares and the stock price will fall precipitously. Similarly, a single cattle ranch can increase its revenue by expanding the size of its livestock herd. The extra cattle will not affect the price of cattle when they are brought to market. But if all ranchers as a group expand their herds, the total output of cattle will increase so much that the price of cattle will decline when the cattle are sold. If the price reduction is relatively large, ranchers as a group might find that their income has fallen despite their having sold a greater number of cattle because the fall in price overwhelms the increase in quantity.

Post Hoc Fallacy You must think very carefully before concluding that because event A precedes event B, A is the cause of B. This kind of faulty reasoning is known as the *post hoc, ergo propter hoc*, or "after this, therefore because of this," fallacy. Noneconomic example: A professional football team hires a new coach and the team's record improves. Is the new coach the cause? Maybe. Perhaps the presence of more experienced and talented players or an easier schedule is the true cause. The rooster crows before dawn but does not cause the sunrise.

Economic example: Many people blamed the Great Depression of the 1930s on the stock market crash of 1929. But the crash did not cause the Great Depression. The same severe weaknesses in the economy that caused the crash caused the Great Depression. The depression would have occurred even without the preceding stock market crash.

Correlation but Not Causation Do not confuse correlation, or connection, with causation. Correlation between two events or two sets of data indicates only that they are associated in some systematic and dependable way. For example, we may find that when variable X increases, Y also increases. But this correlation does not necessarily mean that there is causation—that increases in X cause increases in Y. The relationship could be purely coincidental or dependent on some other factor, Z, not included in the analysis.

Here is an example: Economists have found a positive correlation between education and income. In general, people with more education earn higher incomes than those with less education. Common sense suggests education is the cause and higher incomes are the effect; more education implies a more knowledgeable and productive worker, and such workers receive larger salaries.

But might the relationship be explainable in other ways? Are education and income correlated because the characteristics required for succeeding in education—ability and motivation—are the same ones required to be a productive and highly paid worker? If so, then people with those traits will probably both obtain more education and earn higher incomes. But greater education will not be the sole cause of the higher income.

payoff from this choice for Futureville is greater future production capacity and economic growth. The opportunity cost is fewer consumer goods in the present for Futureville to enjoy.

Is Futureville's choice thus necessarily "better" than Presentville's? That, we cannot say. The different outcomes simply reflect different preferences and priorities in the two countries. But each country will have to live with the economic consequences of its choice.

A Qualification: International Trade

Production possibilities analysis implies that an individual nation is limited to the combinations of output indicated by its production possibilities curve. But we must modify this principle when international specialization and trade exist.

You will see in later chapters that an economy can circumvent, through international specialization and trade, the output limits imposed by its domestic production possibilities curve. Under international specialization and trade, each nation first specializes in the production of those items for which it has the lowest opportunity costs (due to an abundance of the necessary resources). Countries then engage in international trade, with each country exchanging the items that it can produce at the lowest opportunity costs for the items that other countries can produce at the lowest opportunity costs.

International specialization and trade allow a nation to get more of a desired good at less sacrifice of some other good. Rather than sacrifice three units of domestically produced robots to get a third unit of domestically produced pizza, as in Table 1.1, a nation that engages in international specialization and trade might be able to do much better. If it specializes in robots while another country specializes in pizza, then it may be able to obtain the third unit of pizza by trading only two units of domestically produced robots for one unit of foreign-produced pizza. Specialization and trade have the same effect as having more and better resources or discovering improved production techniques; both increase the quantities of capital and consumer goods available to society. Expansion of domestic production possibilities and international trade are two separate routes for obtaining greater output.

QUICK REVIEW 1.4

- Unemployment causes an economy to operate at a point inside its production possibilities curve.
- Increases in resource supplies, improvements in resource quality, and technological advance cause economic growth, which is depicted as an outward shift of the production possibilities curve.
- An economy's present choice of capital and consumer goods helps determine the future location of its production possibilities curve.
- International specialization and trade enable a nation to obtain more goods than its production possibilities curve indicates.

SUMMARY

LO1.1 Define economics and the features of the economic perspective.

Economics is the social science that examines how individuals, institutions, and society make optimal choices under conditions of scarcity. Central to economics is the idea of opportunity cost: the value of the next-best good or service forgone to obtain something.

The economic perspective includes three elements: scarcity and choice, purposeful behavior, and marginal analysis. It sees individuals and institutions making rational decisions based on comparisons of marginal costs and marginal benefits.

LO1.2 Describe the role of economic theory in economics.

Economists employ the scientific method, in which they form and test hypotheses of cause-and-effect relationships to generate theories, laws, and principles. Economists often combine theories into representations called models.

LO1.3 Distinguish microeconomics from macroeconomics and positive economics from normative economics.

Microeconomics examines the decision making of specific economic units or institutions. Macroeconomics looks at the economy as a whole or its major aggregates.

Positive economic analysis deals with facts; normative economics reflects value judgments.

LO1.4 Explain the individual's economizing problem and how trade-offs, opportunity costs, and attainable combinations can be illustrated with budget lines.

Individuals face an economizing problem. Because their wants exceed their incomes, they must decide what to purchase and what to forgo. Society also faces an economizing problem. Societal wants exceed the available resources necessary to fulfill them. Society therefore must decide what to produce and what to forgo.

Graphically, a budget line (or budget constraint) illustrates the economizing problem for individuals. The line shows the various combinations of two products that a consumer can purchase with a specific money income, given the prices of the two products.

LO1.5 List the categories of scarce resources and delineate the nature of society's economizing problem.

Economic resources are inputs into the production process and can be classified as land, labor, capital, or entrepreneurial ability. Economic resources are also known as factors of production or inputs.

Economists illustrate society's economizing problem through production possibilities analysis. Production possibilities tables and curves show the different combinations of goods and services that can be produced in a fully employed economy, assuming that resource quantity, resource quality, and technology are fixed.

LO1.6 Apply production possibilities analysis, increasing opportunity costs, and economic growth.

An economy that is fully employed and thus operating on its production possibilities curve must sacrifice the output of some types of goods and services to increase the production of others. The gain of one type of good or service is always accompanied by an opportunity cost in the form of the loss of some of the other type of good or service.

Because resources are not equally productive in all possible uses, shifting resources from one use to another creates increasing opportunity costs. The production of additional units of one product requires the sacrifice of increasing amounts of the other product.

The optimal (best) point on the production possibilities curve represents the most desirable mix of goods and is determined by expanding the production of each good until its marginal benefit (MB) equals its marginal cost (MC).

LO1.7 Explain how economic growth and international trade increase consumption possibilities.

Over time, technological advances and increases in the quantity and quality of resources enable the economy to produce more of all goods and services, that is, to experience economic growth. Society's choice as to the mix of consumer goods and capital goods in current output is a major determinant of the future location of the production possibilities curve and thus of the extent of economic growth.

International trade enables a nation to obtain more goods from its limited resources than its production possibilities curve indicates.

TERMS AND CONCEPTS

economics
economic perspective
scarcity
opportunity cost
utility
marginal analysis
scientific method
economic principle
other-things-equal assumption
microeconomics
macroeconomics
aggregate
positive economics
normative economics
economizing problem
budget line
economic resources
land
labor
capital
investment
entrepreneurial ability
entrepreneurs
factors of production
consumer goods
capital goods
production possibilities curve
law of increasing opportunity costs
economic growth

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What is an opportunity cost? How does the idea relate to the definition of economics? Which of the following decisions would entail the greater opportunity cost: Allocating a square block in the heart of New York City for a surface parking lot or allocating a square block at the edge of a typical suburb for such a lot? Explain. **LO1.1**
2. Cite three examples of recent decisions that you made in which you, at least implicitly, weighed marginal cost and marginal benefit. **LO1.1**
3. What is meant by the term "utility" and how does the idea relate to purposeful behavior? **LO1.1**
4. What are the key elements of the scientific method and how does this method relate to economic principles and laws? **LO1.2**
5. State (a) a positive economic statement of your choice, and then (b) a normative economic statement relating to your first statement. **LO1.3**
6. How does the slope of a budget line illustrate opportunity costs and trade-offs? How does a budget line illustrate scarcity and the effect of limited incomes? **LO1.4**
7. What are economic resources? What categories do economists use to classify them? Why are resources also called factors of production? Why are they called inputs? **LO1.5**
8. Why is money not considered to be a capital resource in economics? Why is entrepreneurial ability considered a category of economic resource, distinct from labor? What are the major functions of the entrepreneur? **LO1.5**
9. Specify and explain the typical shapes of marginal-benefit and marginal-cost curves. How are these curves used to determine the optimal allocation of resources to a particular product? If current output is such that marginal cost exceeds marginal benefit, should more or fewer resources be allocated to this product? Explain. **LO1.6**
10. Suppose that, on the basis of a nation's production possibilities curve, an economy must sacrifice 10,000 pizzas domestically to get the 1 additional industrial robot it desires but that it can get the robot from another country in exchange for 9,000 pizzas. Relate this information to the following statement: "Through international specialization and trade, a nation can reduce its opportunity cost of obtaining goods and thus 'move outside its production possibilities curve.'" **LO1.7**
11. **LAST WORD** Studies indicate that married men on average earn more income than unmarried men of the same age and education level. Why must we be cautious in concluding that marriage is the cause and higher income is the effect?

REVIEW QUESTIONS

1. Match each term with the correct definition. LO1.1
   economics
   opportunity cost
   marginal analysis
   utility
   a. The next-best thing that must be forgone in order to produce one more unit of a given product.
   b. The pleasure, happiness, or satisfaction obtained from consuming a good or service.
   c. The social science concerned with how individuals, institutions, and society make optimal (best) choices under conditions of scarcity.
   d. Making choices based on comparing marginal benefits with marginal costs.
2. Indicate whether each of the following statements applies to microeconomics or macroeconomics: LO1.3
   a. The unemployment rate in the United States was 8.1 percent in August 2012.
   b. A U.S. software firm discharged 15 workers last month and transferred the work to India.
   c. An unexpected freeze in central Florida reduced the citrus crop and caused the price of oranges to rise.
   d. U.S. output, adjusted for inflation, decreased by 2.4 percent in 2009.
   e. Last week Wells Fargo Bank lowered its interest rate on business loans by one-half of 1 percentage point.
   f. The consumer price index rose by 3.8 percent from August 2011 to August 2012.
3. Suppose that you initially have $100 to spend on books or movie tickets. The books start off costing $25 each and the movie tickets start off costing $10 each. For each of the following situations, would the attainable set of combinations that you can afford increase or decrease? LO1.4
   a. Your budget increases from $100 to $150 while the prices stay the same.
   b. Your budget remains $100, the price of books remains $25, but the price of movie tickets rises to $20.
   c. Your budget remains $100, the price of movie tickets remains $10, but the price of a book falls to $15.
4. Suppose that you are given a $100 budget at work that can be spent only on two items: staplers and pens. If staplers cost $10 each and pens cost $2.50 each, then the opportunity cost of purchasing one stapler is: LO1.4
   a. 10 pens.
   b. 5 pens.
   c. zero pens.
   d. 4 pens.
5. For each of the following situations involving marginal cost (MC) and marginal benefit (MB), indicate whether it would be best to produce more, fewer, or the current number of units. LO1.4
   a. 3,000 units at which $MC = \$10$ and $MB = \$13$.
   b. 11 units at which $MC = \$4$ and $MB = \$3$.
   c. 43,277 units at which $MC = \$99$ and $MB = \$99$.
   d. 82 units at which $MC < MB$.
   e. 5 units at which $MB < MC$.
6. Explain how (if at all) each of the following events affects the location of a country's production possibilities curve: LO1.6
   a. The quality of education increases.
   b. The number of unemployed workers increases.
   c. A new technique improves the efficiency of extracting copper from ore.
   d. A devastating earthquake destroys numerous production facilities.
7. What are the two major ways in which an economy can grow and push out its production possibilities curve? LO1.7
   a. Better weather and nicer cars.
   b. Higher taxes and lower spending.
   c. Increases in resource supplies and advances in technology.
   d. Decreases in scarcity and advances in auditing.

PROBLEMS

1. Potatoes cost Janice $1 per pound, and she has $5.00 that she could possibly spend on potatoes or other items. If she feels that the first pound of potatoes is worth $1.50, the second pound is worth $1.14, the third pound is worth $1.05, and all subsequent pounds are worth $0.30, how many pounds of potatoes will she purchase? What if she only had $2 to spend? LO1.1
2. Pham can work as many or as few hours as she wants at the college bookstore for $9 per hour. But due to her hectic schedule, she has just 15 hours per week that she can spend working at either the bookstore or other potential jobs. One potential job, at a café, will pay her $12 per hour for up to 6 hours per week. She has another job offer at a garage that will pay her $10 an hour for up to 5 hours per week. And she has a potential job at a daycare center that will pay her $8.50 per hour for as many hours as she can work. If her goal is to maximize the amount of money she can make each week, how many hours will she work at the bookstore? LO1.1
3. Suppose you won $15 on a lotto ticket at the local 7-Eleven and decided to spend all the winnings on candy bars and bags of peanuts. Candy bars cost $0.75 each while bags of peanuts cost $1.50 each. LO1.5
   a. Construct a table showing the alternative combinations of the two products that are available.
   b. Plot the data in your table as a budget line in a graph. What is the slope of the budget line? What is the opportunity cost of one more candy bar? Of one more bag of

peanuts? Do these opportunity costs rise, fall, or remain constant as additional units are purchased?

  c. Does the budget line tell you which of the available combinations of candy bars and bags of peanuts to buy?
  d. Suppose that you had won $30 on your ticket, not $15. Show the $30 budget line in your diagram. Has the number of available combinations increased or decreased?

4. Suppose that you are on a desert island and possess exactly 20 coconuts. Your neighbor, Friday, is a fisherman, and he is willing to trade 2 fish for every 1 coconut that you are willing to give him. Another neighbor, Kwame, is also a fisherman, and he is willing to trade 3 fish for every 1 coconut. **LO1.5**
  a. On a single figure, draw budget lines for trading with Friday and for trading with Kwame. (Put coconuts on the vertical axis.)
  b. What is the slope of the budget line from trading with Friday?
  c. What is the slope of the budget line from trading with Kwame?
  d. Which budget line features a larger set of attainable combinations of coconuts and fish?
  e. If you are going to trade coconuts for fish, would you rather trade with Friday or Kwame?

5. To the right is a production possibilities table for consumer goods (automobiles) and capital goods (forklifts): **LO1.6**
  a. Show these data graphically. Upon what specific assumptions is this production possibilities curve based?
  b. If the economy is at point C, what is the cost of one more automobile? Of one more forklift? Which characteristic of the production possibilities curve reflects the law of increasing opportunity costs: its shape or its length?
  c. If the economy characterized by this production possibilities table and curve were producing 3 automobiles and 20 forklifts, what could you conclude about its use of its available resources?
  d. Is production at a point outside the production possibilities curve currently possible? Could a future advance in technology allow production beyond the current production possibilities curve? Could international trade allow a country to consume beyond its current production possibilities curve?

	Production Alternatives				
Type of Production	**A**	**B**	**C**	**D**	**E**
Automobiles	0	2	4	6	8
Forklifts	30	27	21	12	0

6. Look at Figure 1.3. Suppose that the cost of cheese falls, so that the marginal cost of producing pizza decreases. Will the MC curve shift up or down? Will the optimal amount of pizza increase or decrease? **LO1.6**

7. Referring to the table in problem 5, suppose improvement occurs in the technology of producing forklifts but not in the technology of producing automobiles. Draw the new production possibilities curve. Now assume that a technological advance occurs in producing automobiles but not in producing forklifts. Draw the new production possibilities curve. Now draw a production possibilities curve that reflects technological improvement in the production of both goods. **LO1.7**

8. Because investment and capital goods are paid for with savings, higher savings rates reflect a decision to consume fewer goods for the present in order to be able to invest in more goods for the future. Households in China save 40 percent of their annual incomes each year, whereas U.S. households save less than 5 percent. At the same time, production possibilities are growing at roughly 9 percent per year in China but only about 3.5 percent per year in the United States. Use graphical analysis of "present goods" versus "future goods" to explain the difference between China's growth rate and the U.S. growth rate. **LO1.7**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

Graphs and Their Meaning

LO1.8 Understand graphs, curves, and slopes as they relate to economics.

If you glance quickly through this text, you will find many graphs. Some seem simple, while others seem more formidable. All are included to help you visualize and understand economic relationships. Physicists and chemists sometimes illustrate their theories by building arrangements of multicolored wooden balls, representing protons, neutrons, and electrons, that are held in proper relation to one another by wires or sticks. Economists most often use graphs to illustrate their models. By understanding these "pictures," you can more readily comprehend economic relationships.

Construction of a Graph

A *graph* is a visual representation of the relationship between two economic quantities, or variables. The table in Figure 1 is a hypothetical illustration showing the relationship between income and consumption for the economy as a whole. Without even studying economics, we would logically expect that people would buy more goods and services when their incomes go up. Thus, it is not surprising to find in the table that total consumption in the economy increases as total income increases.

The information in the table is expressed graphically in Figure 1. Here is how it is done: We want to show visually how consumption changes as income changes. We therefore represent income on the **horizontal axis** of the graph and consumption on the **vertical axis.**

Now we arrange the vertical and horizontal scales of the graph to reflect the ranges of values of consumption and income and mark the scales in convenient increments. As you can see, the values marked on the scales cover all the values in the table. The increments on both scales are $100.

Because the graph has two dimensions, each point within it represents an income value and its associated consumption value. To find a point that represents one of the five income-consumption combinations in the table in Figure 1, we draw straight lines from the appropriate values on the vertical and horizontal axes. For example, to plot point *c* (the $200 income–$150 consumption point), we draw straight lines up from the horizontal (income) axis at $200 and across from the vertical (consumption) axis at $150. These lines intersect at point *c*, which represents this particular income-consumption combination. You should verify that the other income-consumption combinations shown in the table are properly located in the graph in Figure 1. Finally, by assuming that the same general relationship between income and consumption prevails for all other incomes, we draw a line or smooth curve to connect these points. That line or curve represents the income-consumption relationship.

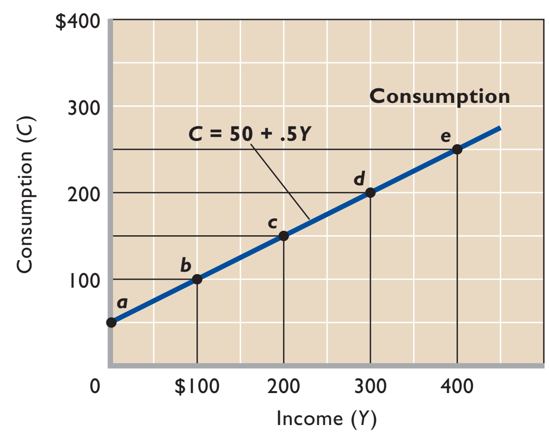
If the curve is a straight line, as in Figure 1, we say the relationship is *linear*. (It is permissible, and even customary, to refer to straight lines in graphs as "curves.")

Direct and Inverse Relationships

The line in Figure 1 slopes upward to the right, so it depicts a direct relationship between income and consumption. By a **direct relationship** (or positive relationship) we mean that two variables—in this case, consumption and

FIGURE 1 Graphing the direct relationship between consumption and income. Two sets of data that are positively or directly related, such as consumption and income, graph as an upsloping line.

Income per Week	Consumption per Week	Point
$ 0	$ 50	*a*
100	100	*b*
200	150	*c*
300	200	*d*
400	250	*e*



income—change in the *same* direction. An increase in consumption is associated with an increase in income; a decrease in consumption accompanies a decrease in income. When two sets of data are positively or directly related, they always graph as an *upsloping* line, as in Figure 1.

In contrast, two sets of data may be inversely related. Consider the table in Figure 2, which shows the relationship between the price of basketball tickets and game attendance at Gigantic State University (GSU). Here we have an **inverse relationship** (or negative relationship) because the two variables change in *opposite* directions. When ticket prices decrease, attendance increases. When ticket prices increase, attendance decreases. The six data points in the table in Figure 2 are plotted in the graph. Observe that an inverse relationship always graphs as a *downsloping* line.

Dependent and Independent Variables

Although it is not always easy, economists seek to determine which variable is the "cause" and which is the "effect." Or, more formally, they seek the independent variable and the dependent variable. The **independent variable** is the cause or source; it is the variable that changes first. The **dependent variable** is the effect or outcome; it is the variable that changes because of the change in the independent variable. As in our income-consumption example, income generally is the independent variable and consumption the dependent variable. Income causes consumption to be what it is rather than the other way around. Similarly, ticket prices (set in advance of the season and printed on the ticket) determine attendance at GSU basketball games; attendance at games does not determine the printed ticket prices for those games. Ticket price is the independent variable and the quantity of tickets purchased is the dependent variable.

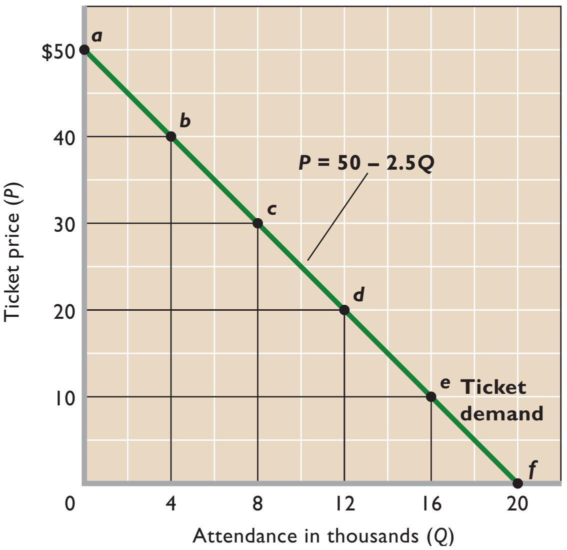
You may recall from your high school courses that mathematicians put the independent variable (cause) on the horizontal axis and the dependent variable (effect) on the vertical axis. Economists are less tidy; their graphing of independent and dependent variables is more arbitrary. Their conventional graphing of the income-consumption relationship is consistent with mathematical convention, but economists put price and cost data on the vertical axis. Hence, economists' graphing of GSU's ticket price–attendance data differs from normal mathematical procedure. This does not present a problem, but we want you to be aware of this fact to avoid any possible confusion.

Other Things Equal

Our simple two-variable graphs purposely ignore many other factors that might affect the amount of consumption occurring at each income level or the number of people who attend GSU basketball games at each possible ticket price. When economists plot the relationship between any two variables, they employ the *ceteris paribus* (other-things-equal) assumption. Thus, in Figure 1 all factors other than income that might affect the amount of consumption are presumed to be constant or unchanged. Similarly, in Figure 2 all factors other than ticket price that might influence attendance at GSU basketball games are assumed constant. In reality, "other things" are not equal; they often change, and when they do, the relationship represented in our two tables and graphs will change. Specifically, the lines we have plotted would *shift* to new locations.

FIGURE 2 Graphing the inverse relationship between ticket prices and game attendance. Two sets of data that are negatively or inversely related, such as ticket price and the attendance at basketball games, graph as a downsloping line.

Ticket Price	Attendance, Thousands	Point
$50	0	*a*
40	4	*b*
30	8	*c*
20	12	*d*
10	16	*e*
0	20	*f*



Consider a stock market "crash." The dramatic drop in the value of stocks might cause people to feel less wealthy and therefore less willing to consume at each level of income. The result might be a downward shift of the consumption line. To see this, you should plot a new consumption line in Figure 1, assuming that consumption is, say, \$20 less at each income level. Note that the relationship remains direct; the line merely shifts downward to reflect less consumption spending at each income level.

Similarly, factors other than ticket prices might affect GSU game attendance. If GSU loses most of its games, attendance at GSU games might be less at each ticket price. To see this, redraw Figure 2 assuming that 2,000 fewer fans attend GSU games at each ticket price.

Slope of a Line

Lines can be described in terms of their slopes. The **slope of a straight line** is the ratio of the vertical change (the rise or drop) to the horizontal change (the run) between any two points of the line.

Positive Slope Between point *b* and point *c* in Figure 1, the rise or vertical change (the change in consumption) is +\$50 and the run or horizontal change (the change in income) is +\$100. Therefore:

$$\text{Slope} = \frac{\text{vertical change}}{\text{horizontal change}} = \frac{+50}{+100} = \frac{1}{2} = .5$$

Note that our slope of $\frac{1}{2}$ or .5 is positive because consumption and income change in the same direction; that is, consumption and income are directly or positively related.

The slope of .5 tells us there will be a \$0.50 increase in consumption for every \$1 increase in income. Similarly, there will be a \$0.50 decrease in consumption for every \$1 decrease in income.

Negative Slope Between any two of the identified points in Figure 2, say, point *c* and point *d*, the vertical change is −10 (the drop) and the horizontal change is +4 (the run). Therefore:

$$\text{Slope} = \frac{\text{vertical change}}{\text{horizontal change}} = \frac{-10}{+4}$$

$$= -2\frac{1}{2} = -2.5$$

This slope is negative because ticket price and attendance have an inverse relationship.

Note that on the horizontal axis attendance is stated in thousands of people. So the slope of −10/+4 or −2.5 means that lowering the price by \$10 will increase attendance by 4,000 people. That ratio also implies that a \$2.50 price reduction will increase attendance by 1,000 persons.

Slopes and Measurement Units The slope of a line will be affected by the choice of units for either variable. If, in our ticket price illustration, we had chosen to measure attendance in individual people, our horizontal change would have been 4,000 and the slope would have been

$$\text{Slope} = \frac{-10}{+4{,}000} = \frac{-1}{+400} = -.0025$$

The slope depends on the way the relevant variables are measured.

Slopes and Marginal Analysis Recall that economics is largely concerned with changes from the status quo. The concept of slope is important in economics because it reflects marginal changes—those involving 1 more (or 1 fewer) unit. For example, in Figure 1 the .5 slope shows that \$0.50 of extra or marginal consumption is associated with each \$1 change in income. In this example, people collectively will consume \$0.50 of any \$1 increase in their incomes and reduce their consumption by \$0.50 for each \$1 decline in income.

Infinite and Zero Slopes Many variables are unrelated or independent of one another. For example, the quantity of wristwatches purchased is not related to the price of bananas. In Figure 3a we represent the price of bananas on the vertical axis and the quantity of watches demanded on the horizontal axis. The graph of their relationship is the line parallel to the vertical axis. The line's vertical slope indicates that the same quantity of watches is purchased no matter what the price of bananas. The slope of vertical lines is *infinite.*

Similarly, aggregate consumption is completely unrelated to the nation's divorce rate. In Figure 3b we put consumption on the vertical axis and the divorce rate on the horizontal axis. The line parallel to the horizontal axis represents this lack of relatedness because the amount of consumption remains the same no matter what happens to the divorce rate. The slope of horizontal lines is *zero*.

Vertical Intercept

A line can be positioned on a graph (without plotting points) if we know just two things: its slope and its vertical intercept. We have already discussed slope. The **vertical intercept** of a line is the point where the line meets the vertical axis. In Figure 1 the intercept is \$50. This intercept means that if current income were zero, consumers would still spend \$50. They might do this through borrowing or by selling some of their assets. Similarly, the \$50 vertical intercept in Figure 2 shows that at a \$50 ticket price, GSU's basketball team would be playing in an empty arena.

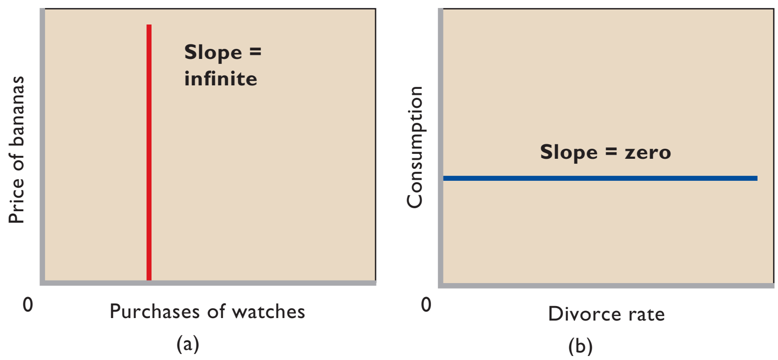


FIGURE 3 **Infinite and zero slopes.** (a) A line parallel to the vertical axis has an infinite slope. Here, purchases of watches remain the same no matter what happens to the price of bananas. (b) A line parallel to the horizontal axis has a slope of zero. In this case, consumption remains the same no matter what happens to the divorce rate. In both (a) and (b), the two variables are totally unrelated to one another.

Equation of a Linear Relationship

If we know the vertical intercept and slope, we can describe a line succinctly in equation form. In its general form, the equation of a straight line is

$$y = a + bx$$

where y = dependent variable
a = vertical intercept
b = slope of line
x = independent variable

For our income-consumption example, if C represents consumption (the dependent variable) and Y represents income (the independent variable), we can write $C = a + bY$. By substituting the known values of the intercept and the slope, we get

$$C = 50 + .5Y$$

This equation also allows us to determine the amount of consumption C at any specific level of income. You should use it to confirm that at the $250 income level, consumption is $175.

When economists reverse mathematical convention by putting the independent variable on the vertical axis and the dependent variable on the horizontal axis, then y stands for the independent variable, rather than the dependent variable in the general form. We noted previously that this case is relevant for our GSU ticket price–attendance data. If P represents the ticket price (independent variable) and Q represents attendance (dependent variable), their relationship is given by

$$P = 50 - 2.5Q$$

where the vertical intercept is 50 and the negative slope is $-2\frac{1}{2}$, or -2.5. Knowing the value of P lets us solve for Q, our dependent variable. You should use this equation to predict GSU ticket sales when the ticket price is $15.

Slope of a Nonlinear Curve

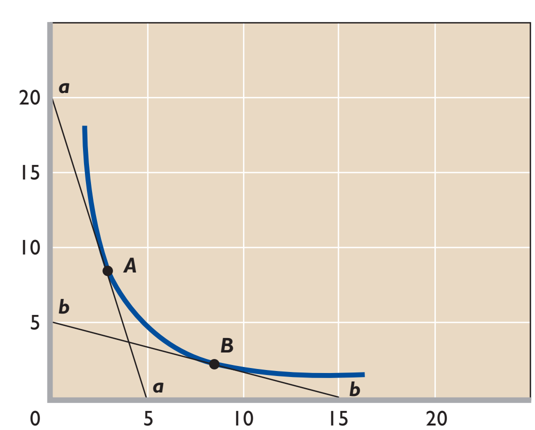
We now move from the simple world of linear relationships (straight lines) to the more complex world of nonlinear relationships (curvy lines). The slope of a straight line is the same at all its points. The slope of a line representing a nonlinear relationship changes from one point to another. Such lines are always referred to as *curves*.

Consider the downsloping curve in Figure 4. Its slope is negative throughout, but the curve flattens as we move down along it. Thus, its slope constantly changes; the curve has a different slope at each point.

To measure the slope at a specific point, we draw a straight line tangent to the curve at that point. A straight line is *tangent* at a point if it touches, but does not intersect, the curve at that point. Thus line *aa* is tangent to the curve in Figure 4 at point A. The slope of the curve at that point is equal to the slope of the tangent line. Specifically, the total vertical change (drop) in the tangent line *aa* is -20 and the total horizontal change (run) is $+5$. Because the slope of the tangent line *aa* is $-20/+5$, or -4, the slope of the curve at point A is also -4.

Line *bb* in Figure 4 is tangent to the curve at point B. Following the same procedure, we find the slope at B to be $-5/+15$, or $-\frac{1}{3}$. Thus, in this flatter part of the curve, the slope is less negative.

FIGURE 4 **Determining the slopes of curves.** The slope of a nonlinear curve changes from point to point on the curve. The slope at any point (say, *B*) can be determined by drawing a straight line that is tangent to that point (line *bb*) and calculating the slope of that line.



APPENDIX SUMMARY

LO1.8 Understand graphs, curves, and slopes as they relate to economics.

Graphs are a convenient and revealing way to represent economic relationships.

Two variables are positively or directly related when their values change in the same direction. The line (curve) representing two directly related variables slopes upward.

Two variables are negatively or inversely related when their values change in opposite directions. The line (curve) representing two inversely related variables slopes downward.

The value of the dependent variable (the "effect") is determined by the value of the independent variable (the "cause").

When the "other factors" that might affect a two-variable relationship are allowed to change, the graph of the relationship will likely shift to a new location.

The slope of a straight line is the ratio of the vertical change to the horizontal change between any two points. The slope of an upsloping line is positive; the slope of a downsloping line is negative.

The slope of a line or curve depends on the units used in measuring the variables. The slope is especially relevant for economics because it measures marginal changes.

The slope of a horizontal line is zero; the slope of a vertical line is infinite.

Together, the vertical intercept and slope of a line determine its location; they are used in expressing the line—and the relationship between the two variables—as an equation.

The slope of a curve at any point is determined by calculating the slope of a straight line tangent to the curve at that point.

APPENDIX TERMS AND CONCEPTS

horizontal axis
vertical axis
direct relationship
inverse relationship
independent variable
dependent variable
slope of a straight line
vertical intercept

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Briefly explain the use of graphs as a way to represent economic relationships. What is an inverse relationship? How does it graph? What is a direct relationship? How does it graph? **LO1.8**
2. Describe the graphical relationship between ticket prices and the number of people choosing to visit amusement parks. Is that relationship consistent with the fact that, historically, park attendance and ticket prices have both risen? Explain. **LO1.8**
3. Look back at Figure 2, which shows the inverse relationship between ticket prices and game attendance at Gigantic State University. (a) Interpret the meaning of both the slope and the intercept. (b) If the slope of the line were steeper, what would that say about the amount by which ticket sales respond to increases in ticket prices? (c) If the slope of the line stayed the same but the intercept increased, what can you say about the amount by which ticket sales respond to increases in ticket prices? **LO1.8**

APPENDIX REVIEW QUESTIONS

1. Indicate whether each of the following relationships is usually a direct relationship or an inverse relationship. **LO1.8**
   - **a.** A sports team's winning percentage and attendance at its home games.
   - **b.** Higher temperatures and sweater sales.
   - **c.** A person's income and how often he or she shops at discount stores.
   - **d.** Higher gasoline prices and miles driven in automobiles.

2. Erin grows pecans. The number of bushels (B) that she can produce depends on the number of inches of rainfall (R) that her orchards get. The relationship is given algebraically as follows: $B = 3{,}000 + 800R$. Match each part of this equation with the correct term. **LO1.8**

B	slope
3,000	dependent variable
800	vertical intercept
R	independent variable

APPENDIX PROBLEMS

1. Graph and label as either direct or indirect the relationships you would expect to find between (a) the number of inches of rainfall per month and the sale of umbrellas, (b) the amount of tuition and the level of enrollment at a university, and (c) the popularity of an entertainer and the price of her concert tickets. **LO1.8**

2. Indicate how each of the following might affect the data shown in the table and graph in Figure 2 of this appendix: **LO1.8**

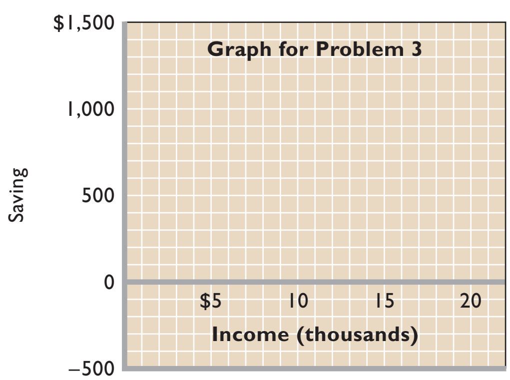
a. GSU's athletic director schedules higher-quality opponents.

b. An NBA team locates in the city where GSU plays.

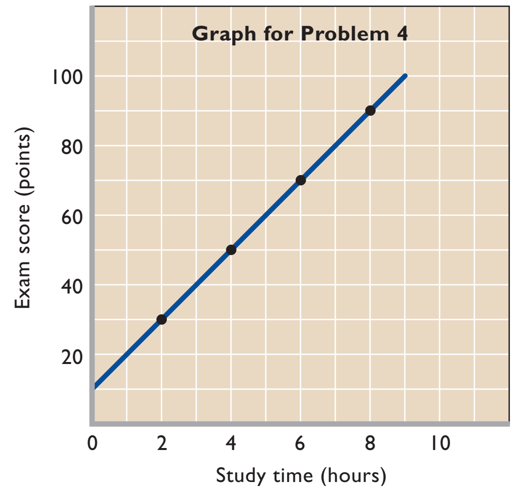
c. GSU contracts to have all its home games televised.

3. The following table contains data on the relationship between saving and income. Rearrange these data into a meaningful order and graph them on the accompanying grid. What is the slope of the line? The vertical intercept? Write the equation that represents this line. What would you predict saving to be at the $12,500 level of income? **LO1.8**

Income per Year	Saving per Year
$15,000	$1,000
0	−500
10,000	500
5,000	0
20,000	1,500



4. Construct a table from the data shown in the accompanying graph. Which is the dependent variable and which is the independent variable? Summarize the data in equation form. **LO1.8**



5. Suppose that when the interest rate on loans is 16 percent, businesses find it unprofitable to invest in machinery and equipment. However, when the interest rate is 14 percent, $5 billion worth of investment is profitable. At 12 percent interest, a total of $10 billion of investment is profitable. Similarly, total investment increases by $5 billion for each successive 2-percentage-point decline in the interest rate. Describe the relevant relationship between the interest rate and investment in a table, on a graph, and as an equation. Put the interest rate on the vertical axis and investment on the horizontal axis. In your equation use the form $i = a + bI$, where i is the interest rate, a is the vertical intercept, b is the slope of the line (which is negative), and I is the level of investment. **LO1.8**

6. Suppose that $C = a + bY$, where C = consumption, a = consumption at zero income, b = slope, and Y = income. **LO1.8**

a. Are C and Y positively related or are they negatively related?

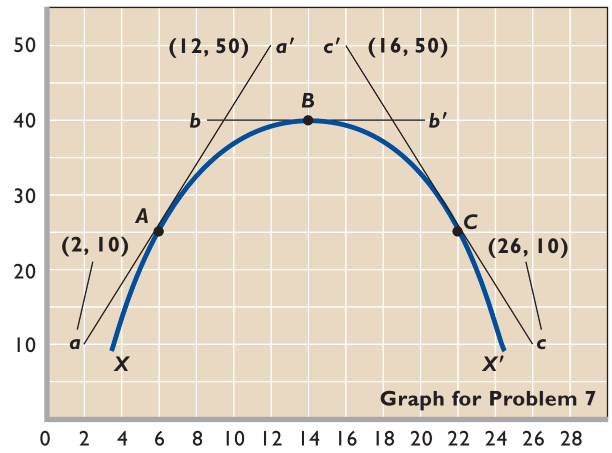
b. If graphed, would the curve for this equation slope upward or slope downward?

c. Are the variables C and Y inversely related or directly related?

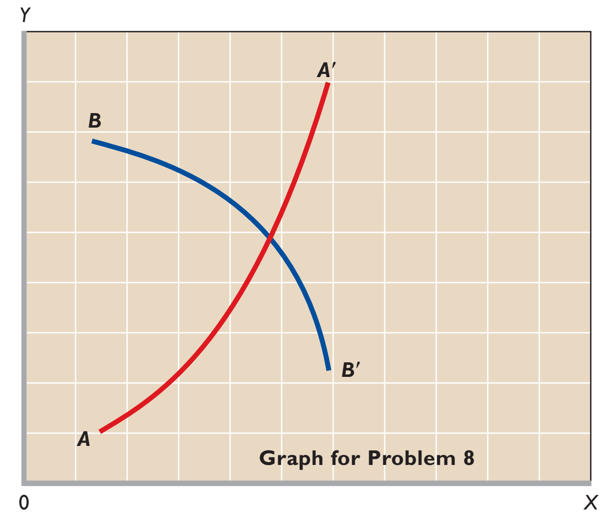
d. What is the value of C if $a = 10$, $b = 0.50$, and $Y = 200$?

e. What is the value of Y if $C = 100$, $a = 10$, and $b = 0.25$?

7. The accompanying graph shows curve XX' and tangents at points A, B, and C. Calculate the slope of the curve at these three points. **LO1.8**



8. In the accompanying graph, is the slope of curve AA' positive or negative? Does the slope increase or decrease as we move along the curve from A to A'? Answer the same two questions for curve BB'. **LO1.8**



CHAPTER 2

The Market System and the Circular Flow

Learning Objectives

LO2.1 **Differentiate between laissez-faire capitalism, the command system, and the market system.**

LO2.2 **List the main characteristics of the market system.**

LO2.3 **Explain how the market system answers the five fundamental questions of what to produce, how to produce, who obtains the output, how to adjust to change, and how to promote progress.**

LO2.4 **Explain the operation of the "invisible hand" and why market economies usually do a better job than command economies at efficiently transforming economic resources into desirable output.**

LO2.5 **Describe the mechanics of the circular flow model.**

LO2.6 **Explain how the market system deals with risk.**

You are at the mall. Suppose you were assigned to compile a list of all the individual goods and services there, including the different brands and variations of each type of product. That task would be daunting and the list would be long! And even though a single shopping mall contains a remarkable quantity and variety of goods, it is only a tiny part of the national economy.

Who decided that the particular goods and services available at the mall and in the broader economy should be produced? How did the producers determine which technology and types of resources to use in producing these particular goods? Who will obtain these products? What accounts for the new and improved products among these goods? This chapter will answer these and related questions.

Economic Systems

LO2.1 Differentiate between laissez-faire capitalism, the command system, and the market system.

Every society needs to develop an **economic system**—a particular set of institutional arrangements and a coordinating mechanism—to respond to the economizing problem. The economic system has to determine what goods are produced, how they are produced, who gets them, how to accommodate change, and how to promote technological progress.

Economic systems differ as to (1) who owns the factors of production and (2) the method used to motivate, coordinate, and direct economic activity.

Economics systems can be classified by the degree to which they rely upon decentralized decision making based upon markets and prices or centralized government control based upon orders and mandates. At one extreme lies *laissez-faire capitalism*, in which government intervention is at a very minimum and markets and prices are allowed to direct nearly all economic activity. At the other extreme lie *command systems*, in which governments have total control over all economic activity. The vast majority of national economies lie somewhere in the middle, utilizing some mixture of centralized government regulation and decentralized markets and prices. These economies are said to have *market systems* or *mixed economies*.

Laissez-Faire Capitalism

In **laissez-faire capitalism**—or "pure capitalism"—the government's role would be limited to protecting private property from theft and aggression and establishing a legal environment in which contracts would be enforced and people could interact in markets to buy and sell goods, services, and resources.

The term "laissez-faire" is the French for "let it be," that is, keep the government from interfering with the economy. Proponents of laissez-faire believe that such interference reduces human welfare. They maintain that any government that intervenes widely in the economy will end up being corrupted by special interests that will use the government's economic influence to benefit themselves rather than society at large.

To prevent that from happening, the proponents of laissez-faire argue that government should restrict itself to preventing individuals and firms from coercing each other. By doing so, it will ensure that only mutually beneficial economic transactions get negotiated and completed. That should lead to the highest possible level of human satisfaction because, after all, who knows better what people want than the people themselves?

It is important to note, however, that no society has ever employed a laissez-faire system. In fact, no government has *ever* limited its economic actions to the short list of functions that would be allowed under laissez-faire. Instead, every government known to history has undertaken a wider range of economic activities, many of which are widely popular and which include industrial safety regulations, various taxes and subsidies, occupational licensing requirements, and income redistribution.

ORIGIN OF THE IDEA

O2.1

Laissez-faire

Thus, you should think of laissez-faire capitalism as a hypothetical system that is viewed by proponents as the ideal to which all economic systems should strive—but which is opposed by those who welcome greater government intervention in the economy.

The Command System

The polar opposite of laissez-faire capitalism is the **command system**, in which government owns most property resources and economic decision making is set by a central economic plan created and enforced by the government. The command system is also known as *socialism* or *communism*.

Under the command system, a central planning board appointed by the government makes all the major decisions concerning the use of resources, the composition and distribution of output, and the organization of production. The government owns most of the business firms, which produce according to government directives. The central planning board determines production goals for each enterprise and specifies the amount of resources to be allocated to each enterprise so that it can reach its production goals. The division of output between capital and consumer goods is centrally decided, and capital goods are allocated among industries on the basis of the central planning board's long-term priorities.

A pure command economy would rely exclusively on a central plan to allocate the government-owned property resources. But, in reality, even the preeminent command economy—the Soviet Union—tolerated some private ownership and incorporated some markets before its collapse in 1992. Recent reforms in Russia and most of the eastern European nations have, to one degree or another, transformed their command economies to capitalistic, market-oriented systems. China's reforms have not gone as far, but they have greatly reduced the reliance on central planning. Although government ownership of resources and capital in China is still extensive, the nation has increasingly relied on free markets to organize and coordinate its economy. North Korea and Cuba are the last prominent remaining examples of largely centrally planned economies. Other countries using mainly the command system include Turkmenistan, Laos, Belarus, Myanmar, and Iran. Later in this chapter, we will explore the main reasons for the general demise of command systems.

The Market System

The vast majority of the world's economies utilize the **market system,** which is also known as *capitalism* or the *mixed economy*.

The market system is characterized by a mixture of centralized government economic initiatives and decentralized actions taken by individuals and firms. The precise mixture varies from country to country, but in each case the system features the private ownership of resources and the use of markets and prices to coordinate and direct economic activity.

In the market system, individuals and businesses seek to achieve their economic goals through their own decisions regarding work, consumption, or production. The system allows for the private ownership of capital, communicates through prices, and coordinates economic activity through markets—places where buyers and sellers come together to buy and sell goods, services, and resources.

Participants act in their own self-interest and goods and services are produced and resources are supplied by whoever is willing and able to do so. The result is competition among independently acting buyers and sellers of each product and resource and an economic system in which decision making is widely dispersed.

The market system also offers high potential monetary rewards that create powerful incentives for existing firms to innovate and for entrepreneurs to pioneer new products and processes despite the financial risks involved and despite most innovations failing to catch on with consumers.

It is the case, however, that in the capitalism practiced in the United States and most other countries, the government plays a substantial role in the economy. It not only provides the rules for economic activity but also promotes economic stability and growth, provides certain goods and services that would otherwise be underproduced or not produced at all, and modifies the distribution of income. The government, however, is not the dominant economic force in deciding what to produce, how to produce it, and who will get it. That force is the market.

Characteristics of the Market System

LO2.2 List the main characteristics of the market system.

An examination of some of the key features of the market system in detail will be very instructive.

Private Property

In a market system, private individuals and firms, not the government, own most of the property resources (land and capital). It is this extensive private ownership of capital that gives capitalism its name. This right of **private property,** coupled with the freedom to negotiate binding legal contracts, enables individuals and businesses to obtain, use, and dispose of property resources as they see fit. The right of property owners to designate who will receive their property when they die helps sustain the institution of private property.

The most important consequence of property rights is that they encourage people to cooperate by helping to ensure that only *mutually agreeable* economic transactions take place. To consider why this is true, imagine a world without legally enforceable property rights. In such a world, the strong could simply take whatever they wanted from the weak without giving them any compensation. But in a world of legally enforceable property rights, any person wanting something from you has to get you to agree to give it to them. And you can say no. The result is that if they really want what you have, they must offer you something that you value more highly in return. That is, they must offer you a mutually agreeable economic transaction—one that benefits you as well as them.

Property rights also encourage investment, innovation, exchange, maintenance of property, and economic growth. Nobody would stock a store, build a factory, or clear land for farming if someone else, or the government itself, could take that property for his or her own benefit.

Property rights also extend to intellectual property through patents, copyrights, and trademarks. Such long-term protection encourages people to write books, music, and computer programs and to invent new products and production processes without fear that others will steal them and the rewards they may bring.

Moreover, property rights facilitate exchange. The title to an automobile or the deed to a cattle ranch assures the buyer that the seller is the legitimate owner. Also, property rights encourage owners to maintain or improve their property so as to preserve or increase its value. Finally, property rights enable people to use their time and resources to produce more goods and services, rather than using them to protect and retain the property they have already produced or acquired.

Freedom of Enterprise and Choice

Closely related to private ownership of property is freedom of enterprise and choice. The market system requires that various economic units make certain choices, which are expressed and implemented in the economy's markets:

- **Freedom of enterprise** ensures that entrepreneurs and private businesses are free to obtain and use economic resources to produce their choice of goods and services and to sell them in their chosen markets.
- **Freedom of choice** enables owners to employ or dispose of their property and money as they see fit. It also allows workers to try to enter any line of work for which they are qualified. Finally, it ensures that consumers are free to buy the goods and services that best satisfy their wants and that their budgets allow.

These choices are free only within broad legal limitations, of course. Illegal choices such as selling human organs or buying illicit drugs are punished through fines and imprisonment. (Global Perspective 2.1 reveals that the degree of economic freedom varies greatly from economy to economy.)

GLOBAL PERSPECTIVE 2.1

Index of Economic Freedom, Selected Economies

The Index of Economic Freedom measures economic freedom using 10 major groupings such as trade policy, property rights, and government intervention, with each category containing more than 50 specific criteria. The index then ranks 179 economies according to their degree of economic freedom. A few selected rankings for 2012 are listed below.

FREE

1 Hong Kong
3 Australia
5 Switzerland

MOSTLY FREE

10 United States
18 Taiwan
26 Germany

MOSTLY UNFREE

99 Brazil
123 India
144 Russia

REPRESSED

158 Argentina
171 Iran
179 North Korea

Source: Used by permission of The Heritage Foundation, **www.heritage.org**.

Self-Interest

In the market system, **self-interest** is the motivating force of the various economic units as they express their free choices. Self-interest simply means that each economic unit tries to achieve its own particular goal, which usually requires delivering something of value to others. Entrepreneurs try to maximize profit or minimize loss. Property owners try to get the highest price for the sale or rent of their resources. Workers try to maximize their utility (satisfaction) by finding jobs that offer the best combination of wages, hours, fringe benefits, and working conditions. Consumers try to obtain the products they want at the lowest possible price and apportion their expenditures to maximize their utility. The motive of self-interest gives direction and consistency to what might otherwise be a chaotic economy.

ORIGIN OF THE IDEA

O2.2
Self-interest

Competition

The market system depends on **competition** among economic units. The basis of this competition is freedom of choice exercised in pursuit of a monetary return. Very broadly defined, competition requires

- Two or more buyers and two or more sellers acting independently in a particular product or resource market. (Usually there are many more than two buyers or sellers.)

- Freedom of sellers and buyers to enter or leave markets, on the basis of their economic self-interest.

Competition among buyers and sellers diffuses economic power within the businesses and households that make up the economy. When there are many buyers and sellers acting independently in a market, no single buyer or seller can dictate the price of the product or resource because others can undercut that price.

Competition also implies that producers can enter or leave an industry; no insurmountable barriers prevent an industry's expanding or contracting. This freedom of an industry to expand or contract provides the economy with the flexibility needed to remain efficient over time. Freedom of entry and exit enables the economy to adjust to changes in consumer tastes, technology, and resource availability.

The diffusion of economic power inherent in competition limits the potential abuse of that power. A producer that charges more than the competitive market price will lose sales to other producers. An employer who pays less than the competitive market wage rate will lose workers to other employers. A firm that fails to exploit new technology will lose profits to firms that do. A firm that produces shoddy products will be punished as customers switch to higher-quality items made by rival firms. Competition is the basic regulatory force in the market system.

Markets and Prices

We may wonder why an economy based on self-interest does not collapse in chaos. If consumers want breakfast cereal, but businesses choose to produce running shoes and resource suppliers decide to make computer software, production would seem to be deadlocked by the apparent inconsistencies of free choices.

In reality, the millions of decisions made by households and businesses are highly coordinated with one another by markets and prices, which are key components of the market system. They give the system its ability to coordinate millions of daily economic decisions. A **market** is an institution or mechanism that brings buyers ("demanders") and sellers ("suppliers") into contact. A market system conveys the decisions made by buyers and sellers of products and resources. The decisions made on each side of the market determine a set of product and resource prices that guide resource owners, entrepreneurs, and consumers as they make and revise their choices and pursue their self-interest.

Just as competition is the regulatory mechanism of the market system, the market system itself is the organizing and coordinating mechanism. It is an elaborate communication network through which innumerable individual free choices are recorded, summarized, and balanced. Those who respond to market signals and heed market dictates are rewarded with greater profit and income; those who do not respond to those signals and choose to ignore market dictates are penalized. Through this mechanism society decides what the economy should produce, how production can be organized efficiently, and how the fruits of production are to be distributed among the various units that make up the economy.

QUICK REVIEW 2.1

- The market system rests on the private ownership of property and on freedom of enterprise and freedom of choice.
- Property rights encourage people to cooperate and make mutually agreeable economic transactions.
- The market system permits consumers, resource suppliers, and businesses to pursue and further their self-interest.
- Competition diffuses economic power and limits the actions of any single seller or buyer.
- The coordinating mechanism of capitalism is a system of markets and prices.

Technology and Capital Goods

In the market system, competition, freedom of choice, self-interest, and personal reward provide the opportunity and motivation for technological advance. The monetary rewards for new products or production techniques accrue directly to the innovator. The market system therefore encourages extensive use and rapid development of complex capital goods: tools, machinery, large-scale factories, and facilities for storage, communication, transportation, and marketing.

Advanced technology and capital goods are important because the most direct methods of production are often the least efficient. The only way to avoid that inefficiency is to rely on capital goods. It would be ridiculous for a farmer to go at production with bare hands. There are huge benefits to be derived from creating and using such capital equipment as plows, tractors, and storage bins. More efficient production means much more abundant output.

Specialization

The extent to which market economies rely on **specialization** is extraordinary. Specialization means using the resources of an individual, firm, region, or nation to produce one or a few goods or services rather than the entire range of goods and services. Those goods and services are then exchanged for a full range of desired products. The majority of consumers produce virtually none of the goods and services they consume, and they consume little or nothing of the items they produce. The person working nine to five installing windows in commercial aircraft may rarely fly. Many farmers sell their milk to the local dairy and then buy margarine at the local grocery store. Society learned long ago that self-sufficiency breeds inefficiency. The jack-of-all-trades may be a very colorful individual but is certainly not an efficient producer.

Division of Labor Human specialization—called the **division of labor**—contributes to a society's output in several ways:

- ***Specialization makes use of differences in ability.*** Specialization enables individuals to take advantage of existing differences in their abilities and skills. If Peyton is strong, athletic, and good at throwing a football and Beyoncé is beautiful, is agile, and can sing, their distribution of talents can be most efficiently used if Peyton plays professional football and Beyoncé records songs and gives concerts.

ORIGIN OF THE IDEA

O2.3

Specialization: division of labor

- ***Specialization fosters learning by doing.*** Even if the abilities of two people are identical, specialization may still be advantageous. By devoting time to a single task, a person is more likely to develop the skills required and to improve techniques than by working at a number of different tasks. You learn to be a good lawyer by studying and practicing law.
- ***Specialization saves time.*** By devoting time to a single task, a person avoids the loss of time incurred in shifting from one job to another. Also, time is saved by not "fumbling around" with tasks that one is not trained to do.

For all these reasons, specialization increases the total output society derives from limited resources.

Geographic Specialization Specialization also works on a regional and international basis. It is conceivable that oranges could be grown in Nebraska, but because of the unsuitability of the land, rainfall, and temperature, the costs would be very high. And it is conceivable that wheat could be grown in Florida, but such production would be costly for similar geographical reasons. So Nebraskans produce products—wheat in particular—for which their resources are best suited, and Floridians do the same, producing oranges and other citrus fruits. By specializing, both economies produce more than is needed locally. Then, very sensibly, Nebraskans and Floridians swap some of their surpluses—wheat for oranges, oranges for wheat.

Similarly, on an international scale, the United States specializes in producing such items as commercial aircraft and software, which it sells abroad in exchange for video cameras from Japan, bananas from Honduras, and woven baskets from Thailand. Both human specialization and geographic specialization are needed to achieve efficiency in the use of limited resources.

Use of Money

A rather obvious characteristic of any economic system is the extensive use of money. Money performs several functions, but first and foremost it is a **medium of exchange.** It makes trade easier.

Specialization requires exchange. Exchange can, and sometimes does, occur through **barter**—swapping goods for goods, say, wheat for oranges. But barter poses serious problems because it requires a *coincidence of wants* between the buyer and the seller. In our example, we assumed that Nebraskans had excess wheat to trade and wanted oranges. And we assumed that Floridians had excess oranges to trade and wanted wheat. So an exchange occurred. But if such a coincidence of wants is missing, trade is stymied.

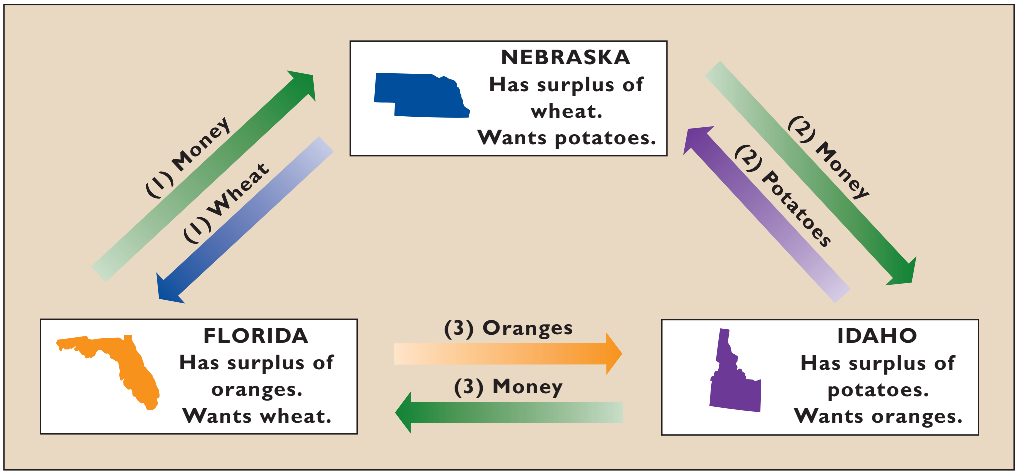
Suppose that Nebraska has no interest in Florida's oranges but wants potatoes from Idaho. And suppose that Idaho wants Florida's oranges but not Nebraska's wheat. And, to complicate matters, suppose that Florida wants some of Nebraska's wheat but none of Idaho's potatoes. We summarize the situation in Figure 2.1.

In none of the cases shown in the figure is there a coincidence of wants. Trade by barter clearly would be difficult. Instead, people in each state use **money,** which is simply a convenient social invention to facilitate exchanges of goods and services. Historically, people have used cattle, cigarettes, shells, stones, pieces of metal, and many other commodities, with varying degrees of success, as money. To serve as money, an item needs to pass only one test: It must be generally acceptable to sellers in exchange for their goods and services. Money is socially defined; whatever society accepts as a medium of exchange *is* money.

Today, most economies use pieces of paper as money. The use of paper dollars (currency) as a medium of exchange is what enables Nebraska, Florida, and Idaho to overcome their trade stalemate, as demonstrated in Figure 2.1.

On a global basis, specialization and exchange are complicated by the fact that different nations have different currencies. But markets in which currencies are bought

FIGURE 2.1 Money facilitates trade when wants do not coincide. The use of money as a medium of exchange permits trade to be accomplished despite a noncoincidence of wants. (1) Nebraska trades the wheat that Florida wants for money from Floridians; (2) Nebraska trades the money it receives from Florida for the potatoes it wants from Idaho; (3) Idaho trades the money it receives from Nebraska for the oranges it wants from Florida.



and sold make it possible for people living in different countries to exchange goods and services without resorting to barter.

Active, but Limited, Government

An active, but limited, government is the final characteristic of market systems in modern advanced industrial economies. Although a market system promotes a high degree of efficiency in the use of its resources, it has certain inherent shortcomings, called "market failures." We will discover in subsequent chapters that governments can often increase the overall effectiveness of a market system in several ways. That being said, governments have their own set of shortcomings that can themselves cause substantial misallocations of resources. Consequently, we will also investigate several types of "government failure."

QUICK REVIEW 2.2

- The market systems of modern industrial economies are characterized by extensive use of technologically advanced capital goods. Such goods help these economies achieve greater efficiency in production.
- Specialization is extensive in market systems; it enhances efficiency and output by enabling individuals, regions, and nations to produce the goods and services for which their resources are best suited.
- The use of money in market systems facilitates the exchange of goods and services that specialization requires.

Five Fundamental Questions

LO2.3 Explain how the market system answers the five fundamental questions of what to produce, how to produce, who obtains the output, how to adjust to change, and how to promote progress.

The key features of the market system help explain how market economies respond to five fundamental questions:

- What goods and services will be produced?
- How will the goods and services be produced?
- Who will get the goods and services?
- How will the system accommodate change?
- How will the system promote progress?

These five questions highlight the economic choices underlying the production possibilities curve discussed in Chapter 1. They reflect the reality of scarce resources in a world of unlimited wants. All economies, whether market or command, must address these five questions.

What Will Be Produced?

How will a market system decide on the specific types and quantities of goods to be produced? The simple answer is this: The goods and services that can be produced at a continuing profit will be produced, while those whose production generates a continuing loss will be discontinued. Profits and losses are the difference between the total revenue (TR) a firm receives from the sale of its products and the total cost (TC) of producing those products. (For economists, total costs include not only wage and salary payments to labor, and interest and rental payments for capital and land, but also payments to the entrepreneur for organizing and combining the other resources to produce a product.)

Continuing economic profit (TR > TC) in an industry results in expanded production and the movement of resources toward that industry. Existing firms grow and new firms enter. The industry expands. Continuing losses (TC > TR) in an industry lead to reduced production and the exit of resources from that industry. Some existing firms shrink in size; others go out of business. The industry contracts. In the market system, consumers are sovereign (in command). **Consumer sovereignty** is crucial in determining the types and quantities of goods produced. Consumers spend their income on the goods they are most willing and able to buy. Through these **"dollar votes"** they register their wants in the market. If the dollar votes for a certain product are great enough to create a profit, businesses will produce that product and offer it for sale. In contrast, if the dollar votes do not create sufficient revenues to cover costs, businesses will not produce the product. So the consumers are sovereign. They collectively direct resources to industries that are meeting consumer wants and away from industries that are not meeting consumer wants.

The dollar votes of consumers determine not only which industries will continue to exist but also which products will survive or fail. Only profitable industries, firms, and products survive. So firms are not as free to produce whatever products they wish as one might otherwise think. Consumers' buying decisions make the production of some products profitable and the production of other products unprofitable, thus restricting the choice of businesses in deciding what to produce. Businesses must match their production choices with consumer choices or else face losses and eventual bankruptcy.

The same holds true for resource suppliers. The employment of resources derives from the sale of the goods and services that the resources help produce. Autoworkers are employed because automobiles are sold. There are few remaining professors of early Latin because there are few people desiring to learn the Latin language. Resource suppliers, desiring to earn income, are not truly free to allocate their resources to the production of goods that consumers do not value highly. Consumers register their preferences in the market; producers and resource suppliers, prompted by their own self-interest, respond appropriately.

CONSIDER THIS . . .

McHits and McMisses

McDonald's has introduced several new menu items over the decades. Some have been profitable "hits," while others have been "misses." In a market system, consumers ultimately decide whether a menu item is profitable and therefore whether it stays on the McDonald's menu.

- Hulaburger (1962)—McMiss
- Filet-O-Fish (1963)—McHit
- Strawberry shortcake (1966)—McMiss
- Big Mac (1968)—McHit
- Hot apple pie (1968)—McHit
- Egg McMuffin (1975)—McHit
- Drive-thru (1975)—McHit
- Chicken McNuggets (1983)—McHit
- Extra Value Meal (1991)—McHit
- McLean Deluxe (1991)—McMiss
- Arch Deluxe (1996)—McMiss
- 55-cent special (1997)—McMiss
- Big Xtra (1999)—McHit
- McSalad Shaker (2000)—McMiss
- McGriddle (2003)—McHit
- Snack Wrap (2006)—McHit

Source: "Polishing the Golden Arches," *Forbes,* June 15, 1998, pp. 42–43, updated. Reprinted by permission of Forbes Media LLC © 2010.

How Will the Goods and Services Be Produced?

What combinations of resources and technologies will be used to produce goods and services? How will the production be organized? The answer: In combinations and ways that minimize the cost per unit of output. This is true because inefficiency drives up costs and lowers profits. As a result, any firm wishing to maximize its profits will make great efforts to minimize production costs. These efforts will include using the right mix of labor and capital, given the prices and productivity of those resources. They also mean locating production facilities optimally to hold down production and transportation expenses.

Those efforts will be intensified if the firm faces competition, as consumers strongly prefer low prices and will shift their purchases over to the firms that can produce a quality product at the lowest possible price. Any firm foolish enough to use higher-cost production methods will go bankrupt as it is undersold by its more efficient competitors who can still make a profit when

TABLE 2.1 **Three Techniques for Producing $15 Worth of Bar Soap**

		Units of Resource					
		Technique 1		Technique 2		Technique 3	
Resource	Price per Unit of Resource	Units	Cost	Units	Cost	Units	Cost
Labor	$2	4	$ 8	2	$ 4	1	$ 2
Land	1	1	1	3	3	4	4
Capital	3	1	3	1	3	2	6
Entrepreneurial ability	3	1	3	1	3	1	3
Total cost of $15 worth of bar soap			$15		$13		$15

selling at a lower price. Simply stated: Competition eliminates high-cost producers.

Least-cost production means that firms must employ the most economically efficient technique of production in producing their output. The most efficient production technique depends on

- The available technology, that is, the available body of knowledge and techniques that can be used to combine economic resources to produce the desired results.
- The prices of the needed resources.

A technique that requires just a few inputs of resources to produce a specific output may be highly inefficient economically if those resources are valued very highly in the market. Economic efficiency requires obtaining a particular output of product with the least input of scarce resources, when both output and resource inputs are measured in dollars and cents. The combination of resources that will produce, say, $15 worth of bathroom soap at the lowest possible cost is the most efficient.

Suppose there are three possible techniques for producing the desired $15 worth of bars of soap. Suppose also that the quantity of each resource required by each production technique and the prices of the required resources are as shown in Table 2.1. By multiplying the required quantities of each resource by its price in each of the three techniques, we can determine the total cost of producing $15 worth of soap by means of each technique.

Technique 2 is economically the most efficient because it is the least costly. It enables society to obtain $15 worth of output by using a smaller amount of resources—$13 worth—than the $15 worth required by the two other techniques. Competition will dictate that producers use technique 2. Thus, the question of how goods will be produced is answered. They will be produced in a least-cost way.

WORKED PROBLEMS

W2.1

Least-cost production

A change in either technology or resource prices, however, may cause a firm to shift from the technology it is using. If the price of labor falls to $0.50, technique 1 becomes more desirable than technique 2. Firms will find they can lower their costs by shifting to a technology that uses more of the resource whose price has fallen. Exercise: Would a new technique involving 1 unit of labor, 4 of land, 1 of capital, and 1 of entrepreneurial ability be preferable to the techniques listed in Table 2.1, assuming the resource prices shown there?

Who Will Get the Output?

The market system enters the picture in two ways when determining the distribution of total output. Generally, any product will be distributed to consumers on the basis of their ability and willingness to pay its existing market price. If the price of some product, say, a small sailboat, is $3,000, then buyers who are willing and able to pay that price will "sail, sail away." Consumers who are unwilling or unable to pay the price will be "sitting on the dock of the bay."

The ability to pay the prices for sailboats and other products depends on the amount of income that consumers have, along with the prices of, and preferences for, various goods. If consumers have sufficient income and want to spend their money on a particular good, they can have it. The amount of income they have depends on (1) the quantities of the property and human resources they supply and (2) the prices those resources command in the resource market. Resource prices (wages, interest, rent, profit) are crucial in determining the size of each person's income and therefore each person's ability to buy part of the economy's output. If a lawyer earning $200 an hour and a janitor earning $10 an hour both work the same number of hours each year, then each year the lawyer will be able to purchase 20 times more of society's output than the janitor.

How Will the System Accommodate Change?

Market systems are dynamic: Consumer preferences, technologies, and resource supplies all change. This means that the particular allocation of resources that is now the most efficient for a specific pattern of consumer tastes, range of technological alternatives, and amount of available resources will become obsolete and inefficient as consumer preferences change, new techniques of production are discovered, and resource supplies change over time. Can the market economy adjust to such changes?

Suppose consumer tastes change. For instance, assume that consumers decide they want more fruit juice and less milk than the economy currently provides. Those changes in consumer tastes will be communicated to producers through an increase in spending on fruit and a decline in spending on milk. Other things equal, prices and profits in the fruit juice industry will rise and those in the milk industry will fall. Self-interest will induce existing competitors to expand output and entice new competitors to enter the prosperous fruit industry and will in time force firms to scale down—or even exit—the depressed milk industry.

The higher prices and greater economic profit in the fruit-juice industry will not only induce that industry to expand but will also give it the revenue needed to obtain the resources essential to its growth. Higher prices and profits will permit fruit producers to attract more resources from less urgent alternative uses. The reverse occurs in the milk industry, where fewer workers and other resources are employed. These adjustments in the economy are appropriate responses to the changes in consumer tastes. This is consumer sovereignty at work.

The market system is a gigantic communications system. Through changes in prices and profits, it communicates changes in such basic matters as consumer tastes and elicits appropriate responses from businesses and resource suppliers. By affecting price and profits, changes in consumer tastes direct the expansion of some industries and the contraction of others. Those adjustments are conveyed to the resource market. As expanding industries employ more resources and contracting industries employ fewer, the resulting changes in resource prices (wages and salaries, for example) and income flows guide resources from the contracting industries to the expanding industries.

This directing or guiding function of prices and profits is a core element of the market system. Without such a system, a government planning board or some other administrative agency would have to direct businesses and resources into the appropriate industries. A similar analysis shows that the system can and does adjust to other fundamental changes—for example, to changes in technology and in the prices of various resources.

How Will the System Promote Progress?

Society desires economic growth (greater output) and higher standards of living (greater output *per person*). How does the market system promote technological improvements and capital accumulation, both of which contribute to a higher standard of living for society?

Technological Advance The market system provides a strong incentive for technological advance and enables better products and processes to supplant inferior ones. An entrepreneur or firm that introduces a popular new product will gain revenue and economic profit at the expense of rivals. Firms that are highly profitable one year may find they are in financial trouble just a few years later. Technological advance also includes new and improved methods that reduce production or distribution costs. By passing part of its cost reduction on to the consumer through a lower product price, a firm can increase sales and obtain economic profit at the expense of rival firms.

Moreover, the market system promotes the *rapid spread* of technological advance throughout an industry. Rival firms must follow the lead of the most innovative firm or else suffer immediate losses and eventual failure. In some cases, the result is **creative destruction:** The creation of new products and production methods completely destroys the market positions of firms that are wedded to existing products and older ways of doing business. Example: The advent of compact discs largely demolished long-play vinyl records, and iPods and other digital technologies subsequently supplanted CDs.

Capital Accumulation Most technological advances require additional capital goods. The market system provides the resources necessary to produce additional capital goods through increased dollar votes for those goods. That is, the market system acknowledges dollar voting for capital goods as well as for consumer goods.

But who counts the dollar votes for capital goods? Answer: Entrepreneurs and business owners. As receivers of profit income, they often use part of that income to purchase capital goods. Doing so yields even greater profit income in the future if the technological innovation that required the additional capital goods is successful. Also, by paying interest or selling ownership shares, the entrepreneur and firm can attract some of the income of households as saving to increase their dollar votes for the production of more capital goods.

QUICK REVIEW 2.3

- The output mix of the market system is determined by profits, which in turn depend heavily on consumer preferences. Economic profits cause industries to expand; losses cause industries to contract.
- Competition forces industries to use the least costly production methods.
- Competitive markets reallocate resources in response to changes in consumer tastes, technological advances, and changes in availability of resources.
- In a market economy, consumer income and product prices determine how output will be distributed.
- Competitive markets create incentives for technological advance and capital accumulation, both of which contribute to increases in standards of living.

The "Invisible Hand"

LO2.4 Explain the operation of the "invisible hand" and why market economies usually do a better job than command economies at efficiently transforming economic resources into desirable output.

In his 1776 book *The Wealth of Nations*, Adam Smith first noted that the operation of a market system creates a curious unity between private interests and social interests. Firms and resource suppliers, seeking to further their own self-interest and operating within the framework of a highly competitive market system, will simultaneously, as though guided by an **"invisible hand,"** promote the public or social interest. For example, we have seen that in a competitive environment, businesses seek to build new and improved products to increase profits. Those enhanced products increase society's well-being. Businesses also use the least costly combination of resources to produce a specific output because doing so is in their self-interest. To act otherwise would be to forgo profit or even to risk business failure. But, at the same time, to use scarce resources in the least costly way is clearly in the social interest as well. It "frees up" resources to produce other things that society desires.

Self-interest, awakened and guided by the competitive market system, is what induces responses appropriate to the changes in society's wants. Businesses seeking to make higher profits and to avoid losses, and resource suppliers pursuing greater monetary rewards, negotiate changes in the allocation of resources and end up with the output that society wants. Competition controls or guides self-interest such that self-interest automatically and quite unintentionally furthers the best interest of society. The invisible hand ensures that when firms maximize their profits and resource suppliers maximize their incomes, these groups also help maximize society's output and income.

Of the various virtues of the market system, three merit reemphasis:

- ***Efficiency*** The market system promotes the efficient use of resources by guiding them into the production of the goods and services most wanted by society. It forces the use of the most efficient techniques in organizing resources for production, and it encourages the development and adoption of new and more efficient production techniques.
- ***Incentives*** The market system encourages skill acquisition, hard work, and innovation. Greater work skills and effort mean greater production and higher incomes, which usually translate into a higher standard of living. Similarly, the assuming of risks by entrepreneurs can result in substantial profit incomes. Successful innovations generate economic rewards.
- ***Freedom*** The major noneconomic argument for the market system is its emphasis on personal freedom. In contrast to central planning, the market system coordinates economic activity without coercion. The market system permits—indeed, it thrives on—freedom of enterprise and choice. Entrepreneurs and workers are free to further their own self-interest, subject to the rewards and penalties imposed by the market system itself.

Of course, no economic system, including the market system, is flawless. In Chapter 4 we will explain two well-known shortcomings of the market system and examine the government policies that try to remedy them.

The Demise of the Command Systems

Our discussion of how a market system answers the five fundamental questions provides insights on why the command systems of the Soviet Union, eastern Europe, and China (prior to its market reforms) failed. Those systems encountered two insurmountable problems.

The Coordination Problem The first difficulty was the coordination problem. The central planners had to coordinate the millions of individual decisions by consumers, resource suppliers, and businesses. Consider the setting up of a factory to produce tractors. The central planners had to establish a realistic annual production target, for example, 1,000 tractors. They then had to make available all the necessary inputs—labor, machinery, electric power, steel, tires, glass, paint, transportation—for the production and delivery of those 1,000 tractors.

Because the outputs of many industries serve as inputs to other industries, the failure of any single industry to achieve its output target caused a chain reaction of repercussions. For example, if iron mines, for want of machinery or labor or transportation, did not supply the steel industry with the required inputs of iron ore, the steel mills were unable to fulfill the input needs of the many industries that depended on steel. Those steel-using industries (such as tractor, automobile, and transportation) were unable to fulfill their planned production goals. Eventually the chain reaction spread to all firms that used steel as an input and from there to other input buyers or final consumers.

The coordination problem became more difficult as the economies expanded. Products and production processes grew more sophisticated and the number of industries requiring planning increased. Planning techniques that worked for the simpler economy proved highly inadequate and inefficient for the larger economy. Bottlenecks and production stoppages became the norm, not the exception. In trying to cope, planners further suppressed product variety, focusing on one or two products in each product category.

A lack of a reliable success indicator added to the coordination problem in the Soviet Union and China prior to its market reforms. We have seen that market economies rely on profit as a success indicator. Profit depends on consumer demand, production efficiency, and product quality. In contrast, the major success indicator for the command economies usually was a quantitative production target that the central planners assigned. Production costs, product quality, and product mix were secondary considerations. Managers and workers often sacrificed product quality and variety because they were being awarded bonuses for meeting quantitative, not qualitative, targets. If meeting production goals meant sloppy assembly work and little product variety, so be it.

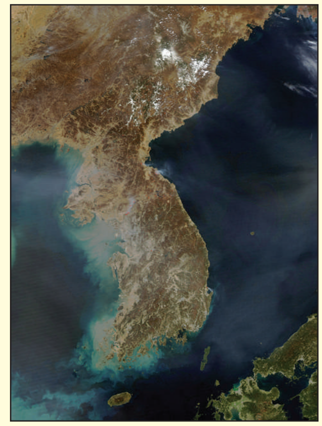
It was difficult at best for planners to assign quantitative production targets without unintentionally producing distortions in output. If the plan specified a production target for producing nails in terms of *weight* (tons of nails), the enterprise made only large nails. But if it specified the target as a *quantity* (thousands of nails), the firm made all small nails, and lots of them! That is precisely what happened in the centrally planned economies.

The Incentive Problem

The command economies also faced an incentive problem. Central planners determined the output mix. When they misjudged how many automobiles, shoes, shirts, and chickens were wanted at the government-determined prices, persistent shortages and surpluses of those products arose. But as long as the managers who oversaw the production of those goods were rewarded for meeting their assigned production goals, they had no incentive to adjust production in response to the shortages and surpluses. And there were no fluctuations in prices and profitability to signal that more or less of certain products was desired. Thus, many products were unavailable or in short supply, while other products were overproduced and sat for months or years in warehouses.

The command systems of the former Soviet Union and China before its market reforms also lacked entrepreneurship. Central planning did not trigger the profit motive, nor did it reward innovation and enterprise. The route for getting ahead was through participation in the political hierarchy of the Communist Party. Moving up the hierarchy meant better housing, better

CONSIDER THIS . . .



The Two Koreas

North Korea is one of the few command economies still standing. After the Second World War, the Korean peninsula was divided into North Korea and South Korea. North Korea, under the influence of the Soviet Union, established a command economy that emphasized government ownership and central government planning. South Korea, protected by the United States, established a market economy based upon private ownership and the profit motive. Today, the differences in the economic outcomes of the two systems are striking:

	North Korea	South Korea
GDP	$40 billion*	$1.6 trillion*
GDP per capita	$1,800*	$32,100*
Exports	$2.5 billion	$556.5 billion
Imports	$3.5 billion	$524.4 billion
Agrculture as % of GDP	23 percent	2.6 percent

*Based on purchasing power equivalencies to the U.S. dollar.

Source: *CIA World Fact Book,* 2011, **www.cia.gov**.

KEY GRAPH

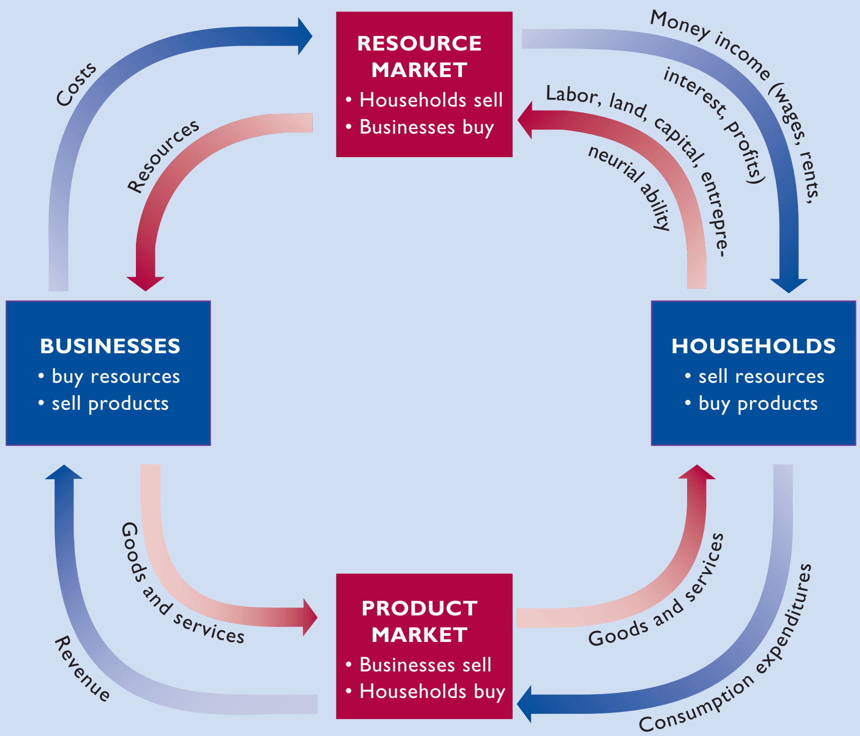


FIGURE 2.2 The circular flow diagram. Resources flow from households to businesses through the resource market, and products flow from businesses to households through the product market. Opposite these real flows are monetary flows. Households receive income from businesses (their costs) through the resource market, and businesses receive revenue from households (their expenditures) through the product market.

QUICK QUIZ FOR FIGURE 2.2

1. The resource market is the place where:
- **a.** households sell products and businesses buy products.
- **b.** businesses sell resources and households sell products.
- **c.** households sell resources and businesses buy resources (or the services of resources).
- **d.** businesses sell resources and households buy resources (or the services of resources).

2. Which of the following would be determined in the product market?
- **a.** a manager's salary.
- **b.** the price of equipment used in a bottling plant.
- **c.** the price of 80 acres of farmland.
- **d.** the price of a new pair of athletic shoes.

3. In this circular flow diagram:
- **a.** money flows counterclockwise.
- **b.** resources flow counterclockwise.
- **c.** goods and services flow clockwise.
- **d.** households are on the selling side of the product market.

4. In this circular flow diagram:
- **a.** households spend income in the product market.
- **b.** firms sell resources to households.
- **c.** households receive income through the product market.
- **d.** households produce goods.

Answers: 1. c; 2. d; 3. b; 4. a

access to health care, and the right to shop in special stores. Meeting production targets and maneuvering through the minefields of party politics were measures of success in "business." But a definition of business success based solely on political savvy was not conducive to technological advance, which is often disruptive to existing products, production methods, and organizational structures.

The Circular Flow Model

LO2.5 Describe the mechanics of the circular flow model.

The dynamic market economy creates continuous, repetitive flows of goods and services, resources, and money. The **circular flow diagram,** shown in **Figure 2.2 (Key Graph),** illustrates those flows for a simplified economy in which there is no government. Observe that in the

ORIGIN OF THE IDEA

O2.4
Circular flow diagram

diagram we group this economy's decision makers into *businesses* and *households*. Additionally, we divide this economy's markets into the *resource market* and the *product market*.

Households

The blue rectangle on the right side of the circular flow diagram in Figure 2.2 represents **households,** which are defined as one or more persons occupying a housing unit. There are currently about 118 million households in the U.S. economy. Households buy the goods and services that businesses make available in the product market. Households obtain the income needed to buy those products by selling resources in the resource market.

All the resources in our no-government economy are ultimately owned or provided by households. For instance, the members of one household or another directly provide all of the labor and entrepreneurial ability in the economy. Households also own all of the land and all of the capital in the economy either directly, as personal property, or indirectly, as a consequence of owning all of the businesses in the economy (and thereby controlling all of the land and capital owned by businesses). Thus, all of the income in the economy—all wages, rents, interest, and profits—flows to households because they provide the economy's labor, land, capital, and entrepreneurial ability.

Businesses

The blue rectangle on the left side of the circular flow diagram represents **businesses,** which are commercial establishments that attempt to earn profits for their owners by offering goods and services for sale. Businesses fall into three main categories.

- A **sole proprietorship** is a business owned and managed by a single person. The proprietor (the owner) may work alone or have employees. Examples include a woman who runs her own tree-cutting business and an independent accountant who, with two assistants, helps his clients with their taxes.
- The **partnership** form of business organization is a natural outgrowth of the sole proprietorship. In a partnership, two or more individuals (the partners) agree to own and operate a business together. They pool their financial resources and business skills to operate the business, and they share any profits or losses that the business may generate. Many law firms and dental practices are organized as partnerships, as are a wide variety of firms in many other industries.
- A **corporation** is an independent legal entity that can—on its own behalf—acquire resources, own assets, produce and sell products, incur debts, extend credit, sue and be sued, and otherwise engage in any legal business activity.

The fact that a corporation is an independent legal entity means that its owners bear no personal financial responsibility for the fulfillment of the corporation's debts and obligations. For instance, if a corporation has failed to repay a loan to a bank, the bank can sue the corporation but not its owners. Professional managers run most corporations. They are hired and supervised by a board of directors that is elected annually by the corporation's owners. Google, Ford, and American Airlines are examples of large corporations, but corporations come in all sizes and operate in every type of industry.

There currently are about 30 million businesses in the United States, ranging from enormous corporations like Walmart, with 2012 sales of $444 billion and 2.2 million employees, to single-person sole proprietorships with sales of less than $100 per day.

Businesses sell goods and services in the product market in order to obtain revenue, and they incur costs in the resource market when they purchase the labor, land, capital, and entrepreneurial ability that they need to produce their respective goods and services.

Product Market

The red rectangle at the bottom of the diagram represents the **product market,** the place where the goods and services produced by businesses are bought and sold. Households use the income they receive from the sale of resources to buy goods and services. The money that they spend on goods and services flows to businesses as revenue.

Resource Market

Finally, the red rectangle at the top of the circular flow diagram represents the **resource market** in which households sell resources to businesses. The households sell resources to generate income, and the businesses buy resources to produce goods and services. Productive resources flow from households to businesses, while money flows from businesses to households in the form of wages, rents, interest, and profits.

To summarize, the circular flow model depicts a complex web of economic activity in which businesses and households are both buyers and sellers. Businesses buy resources and sell products. Households buy products and sell resources. The counterclockwise flow of economic resources and finished products that is illustrated by the red arrows in Figure 2.2 is paid for by the clockwise flow of money income and consumption expenditures illustrated by the blue arrows.

QUICK REVIEW 2.4

- Competition directs individuals and firms to unwittingly promote the social interest, as if guided by a benevolent "invisible hand."
- The command systems of the Soviet Union and prereform China failed as a result of the coordination problem and the incentive problem.
- The circular flow model illustrates how resources flow from households to businesses and how payments for those resources flow from businesses to households.

How the Market System Deals with Risk

LO2.6 Explain how the market system deals with risk.

Producing goods and services is risky. Input shortages can suddenly arise. Consumer preferences can quickly change. Natural disasters can destroy factories and cripple supply chains.

For an economic system to maximize its potential, it must develop methods for assessing and managing risk. The market system does so by confronting business owners with the financial consequences of their decisions. If they manage risks well, they may prosper. If they manage risks poorly, they may lose everything.

The Profit System

As explained in Chapter 1, entrepreneurial ability is the economic resource that organizes and directs the other three resources of land, labor, and capital toward productive uses. The owners of a firm may attempt to supply the entrepreneurial ability themselves. Or they can hire professional managers to supply the necessary leadership and decision making. Either way, it falls to those acting as the firm's entrepreneurs to deal with risk.

They are guided toward sensible decisions by the so-called *profit system*. This system is actually a *profit and loss* system because the entrepreneurs who must deal with risk and uncertainty gain profits if they choose wisely but suffer losses if they choose poorly. That provides them with a large financial incentive to avoid unnecessary risks and make prudent decisions.

By contrast, risk management tends to be done very poorly in command economies because the central planners who must allocate resources and deal with risk do not themselves face the possibility of losing money if they make bad decisions. As government employees, they tend to receive the same salaries whether things go well or poorly.

Shielding Employees and Suppliers from Business Risk

Under the market system, only a firm's owners are subject to business risk and the possibility of losing money. By contrast, the firm's employees and suppliers are shielded from business risk because they are legally entitled to receive their contracted wages and payments on time and in full regardless of whether the firm is earning a profit or generating a loss.

To see how this works, consider a pizza parlor that is being started in a small town. Its investors put up $50,000 to get it going. They rent a storefront, lease ovens, purchase computers, and leave some money set aside as a reserve.

The firm then has to attract employees. To do so, it will offer wage contracts that promise to pay employees every two weeks without regard to whether the firm is making a profit or generating a loss. This guarantee shields the firm's employees from the risks of owning and operating the business. They will get paid even if the pizza parlor is losing money.

In the same way, the contracts that the firm signs with its suppliers and with anyone who loans the firm money (for instance, the local bank) will also specify that they will be paid on time and in full no matter how the firm is doing in terms of profitability.

Dealing with Losses So what happens if the firm starts losing money? In that case, the owners will take the financial hit. To be concrete, suppose that the pizza parlor loses $1,500 during the month of October because it runs up $11,500 in costs but generates only $10,000 in revenue. In that situation, the investors' wealth will shrink by $1,500 as the firm is forced to dip into its reserve to cover the loss. If the firm continues to lose money in subsequent months and exhausts the reserve, the owners will then have to decide whether they want to close the shop or put in additional money in the hope that things will turn around.

CONSIDER THIS . . .

Insurance

Insurance promotes economic growth and investment by transferring risk from those who have a low tolerance for risk to those who have a high tolerance for risk.

Consider fire insurance. Homeowners pay a monthly premium in exchange for which their insurance company guarantees to reimburse them if their house burns down. That guarantee transfers the risk of fire damage from the homeowners (who do not want to bear the risk) to the insurance company's owners (who are happy to bear the risk as a business proposition).

The insurance company will save the premiums that it receives from homeowners to help cover the rebuilding costs of any homes that do end up burning down. But it is quite possible that there will be so many fires that the insurance company will spend more money repairing fire damage than it received in premiums.

If that happens, the insurance company will suffer a loss that will fall upon the insurance company's owners. Their personal wealth will decline in order to make up for the unexpectedly high number of fires. Thus, the insurance company's owners ultimately bear the fire risk; it is to them that the insurance contract transfers risk.

But if the number of fires is unexpectedly small, the insurance company will turn a nice profit and the insurance company's owners will benefit from having been willing to bear the fire risk that the homeowners did not wish to bear.

More importantly, the economy's overall level of investment rises because the availability of insurance means that those who dislike risk are much more willing to invest their savings into the construction and purchase of capital goods like houses, cars, and factories.

But throughout all those months of losses, the suppliers and employees are safeguarded. Because they are paid on time and in full, they are shielded from the firm's business risks and whether it is generating a profit or a loss.

As a result, however, they are not legally entitled to share in the profits if the firm does end up being profitable. That privilege is reserved under the market system for the firm's owners as their reward for bearing business risk. In exchange for making sure that everyone else is shielded if things go badly, the owners are legally entitled to take all of the profits if things go well.

Benefits of Restricting Business Risk to Owners

There are two major benefits that arise from the market system's restriction of business risk to owners and investors.

Attracting Inputs Many people deeply dislike risk and would not be willing to participate in a business venture if they were exposed to the possibility of losing money. That is the case with many workers, who just want to do their jobs and get paid twice a month without having to worry about whether their employer is doing well or not. The same is true for most suppliers, who just want to get paid on time and in full for the inputs they supply to the firm.

For both groups, the concentration of business risk on owners and investors is very welcome because they can supply their resources to a firm without having to worry about the firm's profitability. That sense of security makes it much easier for firms to attract labor and other inputs, which in turn helps the economy innovate and grow.

Focusing Attention The profit system helps to achieve prudent risk management by focusing both the responsibility and the rewards for successfully managing risk onto a firm's owners. They can provide the risk-managing input of entrepreneurial ability themselves or hire it by paying a skilled manager. But either way, some individual's full-time job includes the specialized task of managing risk and making prudent decisions about the allocation of resources. By contrast, in a command system, the responsibility for managing risk tends to be spread out over several layers of government and many different committees so that nobody is personally responsible for bad outcomes.

QUICK REVIEW 2.5

- The market system incentivizes the prudent management of business risk by concentrating any profit or loss upon a firm's owners and investors.
- The market system shields employees, suppliers, and lenders from business risks, but in exchange for that protection, they are excluded from any profit that may be earned.
- By focusing risk on owners and investors, the market system (a) creates an incentive for owners and investors to hire managerial and entrepreneurial specialists to prudently manage business risks and (b) encourages the participation of workers, suppliers, and lenders who dislike risk.

LAST WORD

Shuffling the Deck

Economist Donald Boudreaux Marvels at the Way the Market System Systematically and Purposefully Arranges the World's Tens of Billions of Individual Resources.

In *The Future and Its Enemies*, Virginia Postrel notes the astonishing fact that if you thoroughly shuffle an ordinary deck of 52 playing cards, chances are practically 100 percent that the resulting arrangement of cards has never before existed. *Never.* Every time you shuffle a deck, you produce an arrangement of cards that exists for the first time in history.

The arithmetic works out that way. For a very small number of items, the number of possible arrangements is small. Three items, for example, can be arranged only six different ways. But the number of possible arrangements grows very large very quickly. The number of different ways to arrange five items is 120 . . . for ten items it's 3,628,800 . . . for fifteen items it's 1,307,674,368,000.

The number of different ways to arrange 52 items is 8.066×10^{67}. This is a *big* number. No human can comprehend its enormousness. By way of comparison, the number of possible ways to arrange a mere 20 items is 2,432,902,008,176,640,000—a number larger than the total number of seconds that have elapsed since the beginning of time ten billion years ago—and this number is Lilliputian compared to 8.066×10^{67}.

What's the significance of these facts about numbers? Consider the number of different resources available in the world—my labor, your labor, your land, oil, tungsten, cedar, coffee beans, chickens, rivers, the Empire State Building, [Microsoft] Windows, the wharves at Houston, the classrooms at Oxford, the airport at Miami, and on and on and on. No one can possibly count all of the different productive resources available for our use. But we can be sure that this number is at least in the tens of billions.

When you reflect on how incomprehensibly large is the number of ways to arrange a deck containing a mere 52 cards, the mind boggles at the number of different ways to arrange all the world's resources.

If our world were random—if resources combined together haphazardly, as if a giant took them all into his hands and tossed them down like so many [cards]—it's a virtual certainty that the resulting combination of resources would be useless. Unless this chance arrangement were quickly rearranged according to some productive logic, nothing worthwhile would be produced. We would all starve to death. Because only a tiny fraction of possible arrangements serves human ends, any arrangement will be useless if it is chosen randomly or with inadequate knowledge of how each and every resource might be productively combined with each other.

And yet, we witness all around us an arrangement of resources that's productive and serves human goals. Today's arrangement of resources might not be perfect, but it is vastly superior to most of the trillions upon trillions of other possible arrangements.

How have we managed to get one of the minuscule number of arrangements that works? The answer is private property—a social institution that encourages mutual accommodation.

Private property eliminates the possibility that resource arrangements will be random, for each resource owner chooses a course of action only if it promises rewards to the owner that exceed the rewards promised by all other available courses.

[The result] is a breathtakingly complex and productive arrangement of countless resources. This arrangement emerged over time (and is still emerging) as the result of billions upon billions of individual, daily, small decisions made by people seeking to better employ their resources and labor in ways that other people find helpful.

Source: Abridged from Donald J. Boudreaux, "Mutual Accommodation," *Ideas on Liberty*, May 2000, pp. 4–5. Used by permission of *The Freeman*.

SUMMARY

LO2.1 Differentiate between laissez-faire capitalism, the command system, and the market system.

Laissez-faire capitalism is a hypothetical economic system in which government's role would be restricted to protecting private property and enforcing contracts. All real-world economic systems have featured a more extensive role for government. Governments in command systems own nearly all property and resources and make nearly all decisions about what to produce, how to produce it, and who gets the output. Most countries today, including the United States, have market systems in which the government does play a large role, but in which most property and resources are privately owned and markets are the major force in determining what to produce, how to produce it, and who gets it.

LO2.2 List the main characteristics of the market system.

The market system is characterized by the private ownership of resources, including capital, and the freedom of individuals to engage in economic activities of their choice to advance their material well-being. Self-interest is the driving force of such an economy and competition functions as a regulatory or control mechanism.

In the market system, markets, prices, and profits organize and make effective the many millions of individual economic decisions that occur daily.

Specialization, use of advanced technology, and the extensive use of capital goods are common features of market systems. Functioning as a medium of exchange, money eliminates the problems of bartering and permits easy trade and greater specialization, both domestically and internationally.

LO2.3 Explain how the market system answers the five fundamental questions of what to produce, how to produce, who obtains the output, how to adjust to change, and how to promote progress.

Every economy faces five fundamental questions: (a) What goods and services will be produced? (b) How will the goods and services be produced? (c) Who will get the goods and services? (d) How will the system accommodate change? (e) How will the system promote progress?

The market system produces products whose production and sale yield total revenue sufficient to cover total cost. It does not produce products for which total revenue continuously falls short of total cost. Competition forces firms to use the lowest-cost production techniques.

Economic profit (total revenue minus total cost) indicates that an industry is prosperous and promotes its expansion. Losses signify that an industry is not prosperous and hasten its contraction.

Consumer sovereignty means that both businesses and resource suppliers are subject to the wants of consumers. Through their dollar votes, consumers decide on the composition of output.

The prices that a household receives for the resources it supplies to the economy determine that household's income. This income determines the household's claim on the economy's output. Those who have income to spend get the products produced in the market system.

By communicating changes in consumer tastes to entrepreneurs and resource suppliers, the market system prompts appropriate adjustments in the allocation of the economy's resources. The market system also encourages technological advance and capital accumulation, both of which raise a nation's standard of living.

LO2.4 Explain the operation of the "invisible hand" and why market economies usually do a better job than command economies at efficiently transforming economic resources into desirable output.

Competition, the primary mechanism of control in the market economy, promotes a unity of self-interest and social interests. As if directed by an invisible hand, competition harnesses the self-interested motives of businesses and resource suppliers to further the social interest.

The command systems of the Soviet Union and prereform China met their demise because of coordination difficulties caused by central planning and the lack of a profit incentive. The coordination problem resulted in bottlenecks, inefficiencies, and a focus on a limited number of products. The incentive problem discouraged product improvement, new product development, and entrepreneurship.

LO2.5 Describe the mechanics of the circular flow model.

The circular flow model illustrates the flows of resources and products from households to businesses and from businesses to households, along with the corresponding monetary flows. Businesses are on the buying side of the resource market and the selling side of the product market. Households are on the selling side of the resource market and the buying side of the product market.

LO2.6 Explain how the market system deals with risk.

By focusing business risks onto owners, the market system encourages the participation of workers and suppliers who dislike risk while at the same time creating a strong incentive for owners to manage business risks prudently.

TERMS AND CONCEPTS

economic system	market	"invisible hand"
laissez-faire capitalism	specialization	circular flow diagram
command system	division of labor	households
market system	medium of exchange	businesses
private property	barter	sole proprietorship
freedom of enterprise	money	partnership
freedom of choice	consumer sovereignty	corporation
self-interest	dollar votes	product market
competition	creative destruction	resource market

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Contrast how a market system and a command economy try to cope with economic scarcity. **LO2.1**
2. How does self-interest help achieve society's economic goals? Why is there such a wide variety of desired goods and services in a market system? In what way are entrepreneurs and businesses at the helm of the economy but commanded by consumers? **LO2.2**
3. Why is private property, and the protection of property rights, so critical to the success of the market system? How do property rights encourage cooperation? **LO2.2**
4. What are the advantages of using capital in the production process? What is meant by the term "division of labor"? What are the advantages of specialization in the use of human and material resources? Explain why exchange is the necessary consequence of specialization. **LO2.2**
5. What problem does barter entail? Indicate the economic significance of money as a medium of exchange. What is meant by the statement "We want money only to part with it"? **LO2.2**
6. Evaluate and explain the following statements: **LO2.2**
   a. The market system is a profit-and-loss system.
   b. Competition is the disciplinarian of the market economy.
7. Some large hardware stores, such as Home Depot, boast of carrying as many as 20,000 different products in each store. What motivated the producers of those individual products to make them and offer them for sale? How did the producers decide on the best combinations of resources to use? Who made those resources available, and why? Who decides whether these particular hardware products should continue to be produced and offered for sale? **LO2.3**
8. What is meant by the term "creative destruction"? How does the emergence of MP3 (or iPod) technology relate to this idea? **LO2.3**
9. In a sentence, describe the meaning of the phrase "invisible hand." **LO2.4**
10. In market economies, firms rarely worry about the availability of inputs to produce their products, whereas in command economies input availability is a constant concern. Why the difference? **LO2.4**
11. Distinguish between the resource market and the product market in the circular flow model. In what way are businesses and households both sellers and buyers in this model? What are the flows in the circular flow model? **LO2.5**
12. How does shielding employees and suppliers from business risk help to improve economic outcomes? Who is responsible for managing business risks in the market system? **LO2.6**
13. **LAST WORD** What explains why millions of economic resources tend to get arranged logically and productively rather than haphazardly and unproductively?

REVIEW QUESTIONS

1. Decide whether each of the following descriptions most closely corresponds to being part of a command system, a market system, or a laissez-faire system. **LO2.1**
   a. A woman who wants to start a flower shop finds she cannot do so unless the central government has already decided to allow a flower shop in her area.
   b. Shops stock and sell the goods their customers want but the government levies a sales tax on each transaction in order to fund elementary schools, public libraries, and welfare programs for the poor.
   c. The only taxes levied by the government are to pay for national defense, law enforcement, and a legal system designed to enforce contracts between private citizens.

2. Match each term with the correct definition. LO2.2
   private property
   freedom of enterprise
   mutually agreeable
   freedom of choice
   self-interest
   competition
   market
   a. An institution that brings buyers and sellers together.
   b. The right of private persons and firms to obtain, control, employ, dispose of, and bequeath land, capital, and other property.
   c. The presence in a market of independent buyers and sellers who compete with one another and who are free to enter and exit the market as they each see fit.
   d. The freedom of firms to obtain economic resources, decide what products to produce with those resources, and sell those products in markets of their choice.
   e. What each individual or firm believes is best for itself and seeks to obtain.
   f. Economic transactions willingly undertaken by both the buyer and the seller because each feels that the transaction will make him or her better off.
   g. The freedom of resource owners to dispose of their resources as they think best; of workers to enter any line of work for which they are qualified; and of consumers to spend their incomes in whatever way they feel is most appropriate.
3. True or False: Money must be issued by a government for people to accept it. LO2.2
4. Assume that a business firm finds that its profit is greatest when it produces $40 worth of product A. Suppose also that each of the three techniques shown in the table to the right will produce the desired output. LO2.3
   a. With the resource prices shown, which technique will the firm choose? Why? Will production using that technique entail profit or loss? What will be the amount of that profit or loss? Will the industry expand or contract? When will that expansion or contraction end?
   b. Assume now that a new technique, technique 4, is developed. It combines 2 units of labor, 2 of land, 6 of capital, and 3 of entrepreneurial ability. In view of the resource prices in the table, will the firm adopt the new technique? Explain your answer.
   c. Suppose that an increase in the labor supply causes the price of labor to fall to $1.50 per unit, all other resource prices remaining unchanged. Which technique will the producer now choose? Explain.
   d. "The market system causes the economy to conserve most in the use of resources that are particularly scarce in supply. Resources that are scarcest relative to the demand for them have the highest prices. As a result, producers use these resources as sparingly as is possible." Evaluate this statement. Does your answer to part *c*, above, bear out this contention? Explain.

		Resource Units Required		
Resource	Price per Unit of Resource	Technique 1	Technique 2	Technique 3
Labor	$3	5	2	3
Land	4	2	4	2
Capital	2	2	4	5
Entrepreneurial ability	2	4	2	4

5. Identify each of the following quotes as being an example of either: the coordination problem, the invisible hand, creative destruction, or the incentive problem. LO2.4
   a. "If you compare a list of today's most powerful and profitable companies with a similar list from 30 years ago, you will see lots of new entries."
   b. "Managers in the old Soviet Union often sacrificed product quality and variety because they were being awarded bonuses for quantitative, not qualitative, targets."
   c. "Each day, central planners in the old Soviet Union were tasked with setting 27 million prices—correctly."
   d. "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest."
6. True or False: Households sell finished products to businesses. LO2.6
7. Franklin, John, Henry, and Harry have decided to pool their financial resources and business skills in order to open up and run a coffee shop. They will share any profits or losses that the business generates and will be personally responsible for making good on any debt that their business undertakes. Their business should be classified as a: LO2.6
   a. Corporation.
   b. Sole proprietorship.
   c. Partnership.
   d. None of the above.
8. Ted and Fred are the owners of a gas station. They invested $150,000 each and pay an employee named Lawrence $35,000 per year. This year revenues are $900,000, while costs are $940,000. Who is legally responsible for bearing the $40,000 loss? LO2.6
   a. Lawrence.
   b. Ted.
   c. Fred.
   d. Ted and Fred.
   c. Lawrence, Ted, and Fred.

PROBLEMS

1. Table 2.1 contains information on three techniques for producing $15 worth of bar soap. Assume that we said "$15 worth of bar soap" because soap cost $3 per bar and all three techniques produce 5 bars of soap ($15 = $3 per bar × 5 bars). So you know each technique produces 5 bars of soap. **LO2.3**
   - **a.** What technique will you want to use if the price of a bar of soap falls to $2.75? What if the price of a bar of soap rises to $4? To $5?
   - **b.** How many bars of soap will you want to produce if the price of a bar of soap falls to $2.00?
   - **c.** Suppose that the price of soap is again $3 per bar but that the prices of all four resources are now $1 per unit. Which is now the least-profitable technique?
   - **d.** If the resource prices return to their original levels (the ones shown in the table), but a new technique is invented that can produce 3 bars of soap (yes, 3 bars, not 5 bars!), using 1 unit of each of the four resources, will firms prefer the new technique?
2. Suppose Natasha currently makes $50,000 per year working as a manager at a cable TV company. She then develops two possible entrepreneurial business opportunities. In one, she will quit her job to start an organic soap company. In the other, she will try to develop an Internet-based competitor to the local cable company. For the soap-making opportunity, she anticipates annual revenue of $465,000 and costs for the necessary land, labor, and capital of $395,000 per year. For the Internet opportunity, she anticipates costs for land, labor, and capital of $3,250,000 per year as compared to revenues of $3,275,000 per year. (a) Should she quit her current job to become an entrepreneur? (b) If she does quit her current job, which opportunity would she pursue? **LO2.3**
3. With current technology, suppose a firm is producing 400 loaves of banana bread daily. Also assume that the least-cost combination of resources in producing those loaves is 5 units of labor, 7 units of land, 2 units of capital, and 1 unit of entrepreneurial ability, selling at prices of $40, $60, $60, and $20, respectively. If the firm can sell these 400 loaves at $2 per unit, what is its total revenue? Its total cost? Its profit or loss? Will it continue to produce banana bread? If this firm's situation is typical for the other makers of banana bread, will resources flow toward or away from this bakery good? **LO2.3**
4. Let's put dollar amounts on the flows in the circular flow diagram of Figure 2.2. **LO2.5**
   - **a.** Suppose that businesses buy a total of $100 billion of the four resources (labor, land, capital, and entrepreneurial ability) from households. If households receive $60 billion in wages, $10 billion in rent, and $20 billion in interest, how much are households paid for providing entrepreneurial ability?
   - **b.** If households spend $55 billion on goods and $45 billion on services, how much in revenues do businesses receive in the product market?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **TWO**

PRICE, QUANTITY, AND EFFICIENCY

CHAPTER **3**

Demand, Supply, and Market Equilibrium

Learning Objectives

LO3.1 Characterize and give examples of markets.

LO3.2 Describe *demand* and explain how it can change.

LO3.3 Describe *supply* and explain how it can change.

LO3.4 Relate how supply and demand interact to determine market equilibrium.

LO3.5 Explain how changes in supply and demand affect equilibrium prices and quantities.

LO3.6 Identify what government-set prices are and how they can cause product surpluses and shortages.

LO3.7 (Appendix) Illustrate how supply and demand analysis can provide insights on actual-economy situations.

ORIGIN OF THE IDEA

O3.1

Demand and supply

The model of supply and demand is the economics profession's greatest contribution to human understanding because it explains the operation of the markets on which we depend for nearly everything that we eat, drink, or consume. The model is so powerful and so widely used that to many people it *is* economics.

This chapter explains how the model works and how it can explain both the *quantities* that are bought and sold in markets as well as the *prices* at which they trade.

Markets

LO3.1 Characterize and give examples of markets.

Markets bring together buyers ("demanders") and sellers ("suppliers"). The corner gas station, an e-commerce site, the local music store, a farmer's roadside stand—all are familiar markets. The New York Stock Exchange and the Chicago Board of Trade are markets in which buyers and sellers from all over the world communicate with one another to buy and sell bonds, stocks, and commodities. Auctioneers bring together potential buyers and sellers of art, livestock, used farm equipment, and, sometimes, real estate. In labor markets, new college graduates "sell" and employers "buy" specific labor services.

Some markets are local; others are national or international. Some are highly personal, involving face-to-face contact between demander and supplier; others are faceless, with buyer and seller never seeing or knowing each other.

To keep things simple, we will focus in this chapter on markets in which large numbers of independently acting buyers and sellers come together to buy and sell standardized products. Markets with these characteristics are the economy's most highly competitive. They include the wheat market, the stock market, and the market for foreign currencies. All such markets involve demand, supply, price, and quantity. As you will soon see, the price is "discovered" through the interacting decisions of buyers and sellers.

Demand

LO3.2 Describe *demand* and explain how it can change.

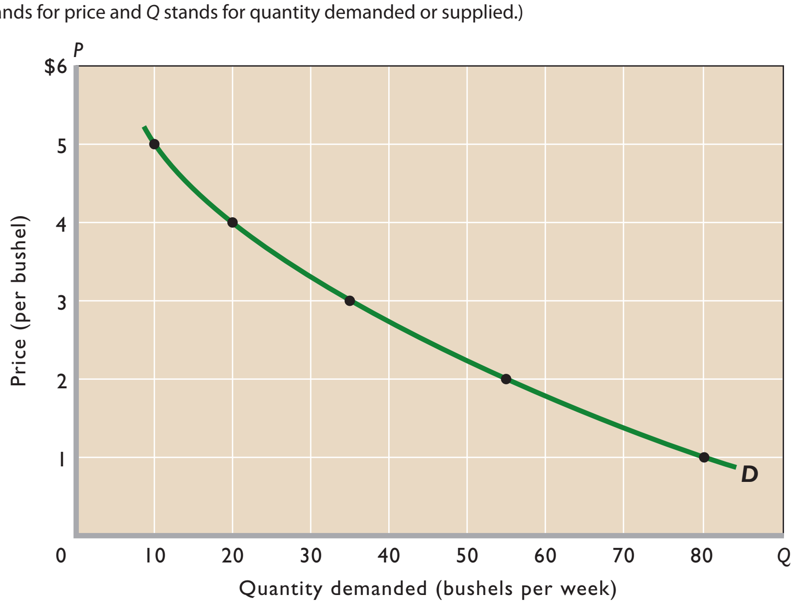
Demand is a schedule or a curve that shows the various amounts of a product that consumers are willing and able to purchase at each of a series of possible prices during a specified period of time.[1] Demand shows the quantities of a product that will be purchased at various possible prices, *other things equal*. Demand can easily be shown in table form. The table in Figure 3.1 is a hypothetical **demand schedule** for a *single consumer* purchasing bushels of corn.

The table reveals the relationship between the various prices of corn and the quantity of corn a particular consumer would be willing and able to purchase at each of these prices. We say "willing and able" because willingness alone is not effective in the market. You may be willing to buy a plasma television set, but if that willingness is not backed by the necessary dollars, it will not be effective and, therefore, will not be reflected in the market. In the table in Figure 3.1, if the price of corn were $5 per bushel, our consumer would be willing and able to buy 10 bushels per week; if it were $4, the consumer would be willing and able to buy 20 bushels per week; and so forth.

[1]This definition obviously is worded to apply to product markets. To adjust it to apply to resource markets, substitute the word "resource" for "product" and the word "businesses" for "consumers."

FIGURE 3.1 An individual buyer's demand for corn. Because price and quantity demanded are inversely related, an individual's demand schedule graphs as a downsloping curve such as *D*. Other things equal, consumers will buy more of a product as its price declines and less of the product as its price rises. (Here and in later figures, *P* stands for price and *Q* stands for quantity demanded or supplied.)

Demand for Corn	
Price per Bushel	**Quantity Demanded per Week**
$5	10
4	20
3	35
2	55
1	80



The table does not tell us which of the five possible prices will actually exist in the corn market. That depends on the interaction between demand and supply. Demand is simply a statement of a buyer's plans, or intentions, with respect to the purchase of a product.

To be meaningful, the quantities demanded at each price must relate to a specific period—a day, a week, a month. Saying "A consumer will buy 10 bushels of corn at $5 per bushel" is meaningless. Saying "A consumer will buy 10 bushels of corn *per week* at $5 per bushel" is meaningful. Unless a specific time period is stated, we do not know whether the demand for a product is large or small.

Law of Demand

A fundamental characteristic of demand is this: Other things equal, as price falls, the quantity demanded rises, and as price rises, the quantity demanded falls. In short, there is a negative or *inverse* relationship between price and quantity demanded. Economists call this inverse relationship the **law of demand.**

ORIGIN OF THE IDEA

O3.2

Law of demand

The other-things-equal assumption is critical here. Many factors other than the price of the product being considered affect the amount purchased. For example, the quantity of Nikes purchased will depend not only on the price of Nikes but also on the prices of such substitutes as Reeboks, Adidas, and New Balances. The law of demand in this case says that fewer Nikes will be purchased if the price of Nikes rises and if the prices of Reeboks, Adidas, and New Balances all remain constant. In short, if the *relative price* of Nikes rises, fewer Nikes will be bought. However, if the price of Nikes and the prices of all other competing shoes increase by some amount—say, $5—consumers might buy more, fewer, or the same number of Nikes.

Why the inverse relationship between price and quantity demanded? Let's look at three explanations, beginning with the simplest one:

- The law of demand is consistent with common sense. People ordinarily *do* buy more of a product at a low price than at a high price. Price is an obstacle that deters consumers from buying. The higher that obstacle, the less of a product they will buy; the lower the price obstacle, the more they will buy. The fact that businesses have "sales" to clear out unsold items is evidence of their belief in the law of demand.
- In any specific time period, each buyer of a product will derive less satisfaction (or benefit, or utility) from each successive unit of the product consumed. The second Big Mac will yield less satisfaction to the consumer than the first, and the third still less than the second. That is, consumption is subject to **diminishing marginal utility.** And because successive units of a particular product yield less and less marginal utility, consumers will buy additional units only if the price of those units is progressively reduced.

ORIGIN OF THE IDEA

O3.3

Diminishing marginal utility

- We can also explain the law of demand in terms of income and substitution effects. The **income effect** indicates that a lower price increases the purchasing power of a buyer's money income, enabling the buyer to purchase more of the product than before. A higher price has the opposite effect. The **substitution effect** suggests that at a lower price buyers have the incentive to substitute what is now a less expensive product for other products that are now *relatively* more expensive. The product whose price has fallen is now "a better deal" relative to the other products.

For example, a decline in the price of chicken will increase the purchasing power of consumer incomes, enabling people to buy more chicken (the income effect). At a lower price, chicken is relatively more attractive and consumers tend to substitute it for pork, lamb, beef, and fish (the substitution effect). The income and substitution effects combine to make consumers able and willing to buy more of a product at a low price than at a high price.

ORIGIN OF THE IDEA

O3.4

Income and substitution effects

The Demand Curve

The inverse relationship between price and quantity demanded for any product can be represented on a simple graph, in which, by convention, we measure *quantity demanded* on the horizontal axis and *price* on the vertical axis. In the graph in Figure 3.1 we have plotted the five price-quantity data points listed in the accompanying table and connected the points with a smooth curve, labeled *D*. Such a curve is called a **demand curve.** Its downward slope reflects the law of demand—people buy more of a product, service, or resource as its price falls. The relationship between price and quantity demanded is inverse (or negative).

The table and graph in Figure 3.1 contain exactly the same data and reflect the same relationship between price and quantity demanded. But the graph shows that relationship much more simply and clearly than a table or a description in words.

Market Demand

So far, we have concentrated on just one consumer. But competition requires that more than one buyer be present in each market. By adding the quantities demanded by all consumers at each of the various possible prices, we can get from *individual* demand to *market* demand. If there are just three buyers in the market, as represented in the table in Figure 3.2, it is relatively easy to determine the total quantity demanded at each price. Figure 3.2 shows the graphical summing procedure: At each price we sum horizontally the quantities demanded by Joe, Jen, and Jay to obtain the total quantity demanded at that price; we then plot the price and the total quantity demanded as one point on the market demand curve. At the price of \$3, for example, the three individual curves yield a total quantity demanded of 100 bushels (= 35 + 39 + 26).

Competition, of course, ordinarily entails many more than three buyers of a product. To avoid hundreds or thousands or millions of additions, we suppose that all the buyers in a market are willing and able to buy the same amounts at each of the possible prices. Then we just multiply those amounts by the number of buyers to obtain the market demand. That is how we arrived at the demand schedule and demand curve D_1 in Figure 3.3 for a market of 200 corn buyers, each with a demand as shown in the table in Figure 3.1.

In constructing a demand curve such as D_1 in Figure 3.3, economists assume that price is the most important influence on the amount of any product purchased. But economists know that other factors can and do affect purchases. These factors, called **determinants of demand,** are assumed to be constant when a demand curve like D_1 is drawn. They are the "other things equal" in the relationship between price and quantity demanded. When any of these determinants changes, the demand curve will shift to the right or left. For this reason, determinants of demand are sometimes referred to as *demand shifters*.

The basic determinants of demand are (1) consumers' tastes (preferences), (2) the number of buyers in the market, (3) consumers' incomes, (4) the prices of related goods, and (5) consumer expectations.

Changes in Demand

A change in one or more of the determinants of demand will change the demand data (the demand schedule) in the table accompanying Figure 3.3 and therefore the location of the demand curve there. A change in the

FIGURE 3.2 Market demand for corn, three buyers. The market demand curve *D* is the horizontal summation of the individual demand curves (D_1, D_2, and D_3) of all the consumers in the market. At the price of \$3, for example, the three individual curves yield a total quantity demanded of 100 bushels (= 35 + 39 + 26).

Market Demand for Corn, Three Buyers

Price per Bushel	Quantity Demanded: Joe		Jen		Jay		Total Quantity Demanded per Week
\$5	10	+	12	+	8	=	30
4	20	+	23	+	17	=	60
3	35	+	39	+	26	=	100
2	55	+	60	+	39	=	154
1	80	+	87	+	54	=	221

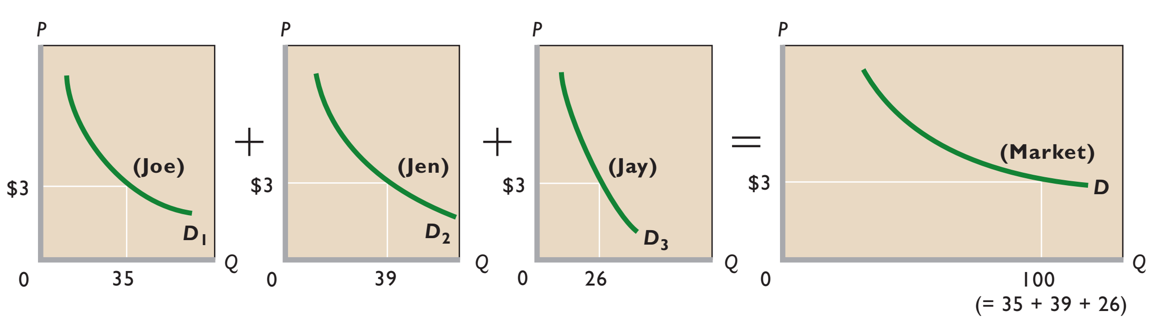
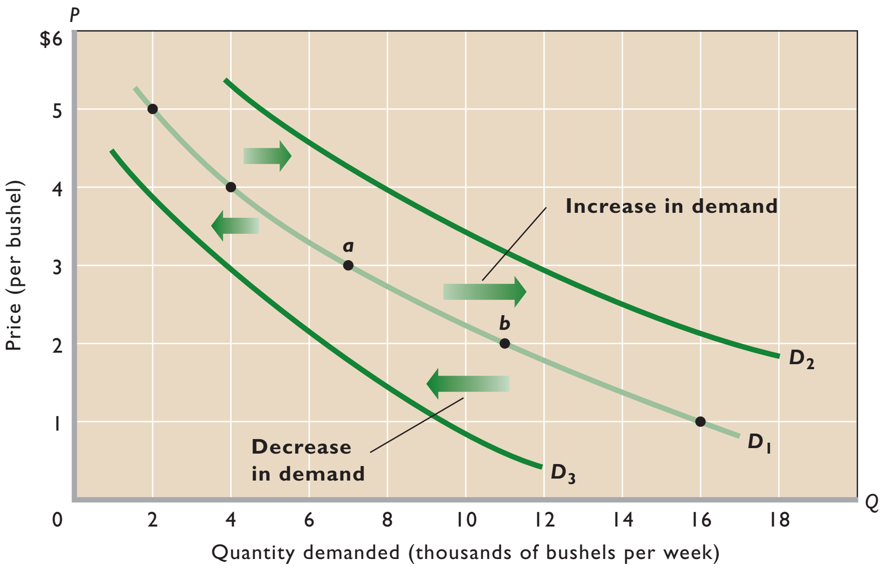


FIGURE 3.3 Changes in the demand for corn. A change in one or more of the determinants of demand causes a change in demand. An increase in demand is shown as a shift of the demand curve to the right, as from D_1 to D_2. A decrease in demand is shown as a shift of the demand curve to the left, as from D_1 to D_3. These changes in demand are to be distinguished from a change in quantity demanded, which is caused by a change in the price of the product, as shown by a movement from, say, point *a* to point *b* on fixed demand curve D_1.



Market Demand for Corn, 200 Buyers, (D_1)

(1) Price per Bushel	(2) Total Quantity Demanded per Week
$5	2,000
4	4,000
3	7,000
2	11,000
1	16,000

demand schedule or, graphically, a shift in the demand curve is called a *change in demand*.

If consumers desire to buy more corn at each possible price than is reflected in column 2 in the table in Figure 3.3, that *increase in demand* is shown as a shift of the demand curve to the right, say, from D_1 to D_2. Conversely, a *decrease in demand* occurs when consumers buy less corn at each possible price than is indicated in column 2. The leftward shift of the demand curve from D_1 to D_3 in Figure 3.3 shows that situation.

Now let's see how changes in each determinant affect demand.

Tastes A favorable change in consumer tastes (preferences) for a product—a change that makes the product more desirable—means that more of it will be demanded at each price. Demand will increase; the demand curve will shift rightward. An unfavorable change in consumer preferences will decrease demand, shifting the demand curve to the left.

New products may affect consumer tastes; for example, the introduction of digital cameras greatly decreased the demand for film cameras. Consumers' concern over the health hazards of cholesterol and obesity have increased the demand for broccoli, low-calorie beverages, and fresh fruit while decreasing the demand for beef, veal, eggs, and whole milk. Over the past several years, the demand for coffee drinks and table wine has greatly increased, driven by a change in tastes. So, too, has the demand for touch-screen mobile phones and fuel-efficient hybrid vehicles.

Number of Buyers An increase in the number of buyers in a market is likely to increase demand; a decrease in the number of buyers will probably decrease demand. For example, the rising number of older persons in the United States in recent years has increased the demand for motor homes, medical care, and retirement communities. Large-scale immigration from Mexico has greatly increased the demand for a range of goods and services in the Southwest, including Mexican food products in local grocery stores. Improvements in communications have given financial markets international range and have thus increased the demand for stocks and bonds. International trade agreements have reduced foreign trade barriers to American farm commodities, increasing the number of buyers and therefore the demand for those products.

In contrast, emigration (out-migration) from many small rural communities has reduced the population and thus the demand for housing, home appliances, and auto repair in those towns.

Income How changes in income affect demand is a more complex matter. For most products, a rise in income causes an increase in demand. Consumers typically buy more steaks, furniture, and electronic equipment as their incomes increase. Conversely, the demand for such products declines as their incomes fall. Products whose demand varies *directly* with money income are called *superior goods*, or **normal goods.**

Although most products are normal goods, there are some exceptions. As incomes increase beyond some point, the demand for used clothing, retread tires, and third-hand automobiles may decrease because the higher incomes enable consumers to buy new versions of those products. Rising incomes may also decrease the demand for soy-enhanced hamburger. Similarly, rising incomes may cause the demand for charcoal grills to decline as wealthier consumers switch to gas grills. Goods whose demand varies *inversely* with money income are called **inferior goods.**

Prices of Related Goods A change in the price of a related good may either increase or decrease the demand for a product, depending on whether the related good is a substitute or a complement:

- A **substitute good** is one that can be used in place of another good.
- A **complementary good** is one that is used together with another good.

Substitutes Häagen-Dazs ice cream and Ben & Jerry's ice cream are substitute goods or, simply, *substitutes.* When two products are substitutes, an increase in the price of one will increase the demand for the other. Conversely, a decrease in the price of one will decrease the demand for the other. For example, when the price of Häagen-Dazs ice cream rises, consumers will buy less of it and increase their demand for Ben & Jerry's ice cream. When the price of Colgate toothpaste declines, the demand for Crest decreases. So it is with other product pairs such as Nikes and Reeboks, Budweiser and Miller beer, or Chevrolets and Fords. They are *substitutes in consumption*.

Complements Because complementary goods (or, simply, *complements*) are used together, they are typically demanded jointly. Examples include computers and software, cell phones and cellular service, and snowboards and lift tickets. If the price of a complement (for example, lettuce) goes up, the demand for the related good (salad dressing) will decline. Conversely, if the price of a complement (for example, tuition) falls, the demand for a related good (textbooks) will increase.

Unrelated Goods The vast majority of goods are not related to one another and are called *independent goods*. Examples are butter and golf balls, potatoes and automobiles, and bananas and wristwatches. A change in the price of one has little or no effect on the demand for the other.

Consumer Expectations Changes in consumer expectations may shift demand. A newly formed expectation of higher future prices may cause consumers to buy now in order to "beat" the anticipated price rises, thus increasing current demand. That is often what happens in so-called hot real estate markets. Buyers rush in because they think the price of new homes will continue to escalate rapidly. Some buyers fear being "priced out of the market" and therefore not obtaining the home they desire. Other buyers—speculators—believe they will be able to sell the houses later at a higher price. Whichever their motivation, these buyers increase the current demand for houses.

Similarly, a change in expectations concerning future income may prompt consumers to change their current spending. For example, first-round NFL draft choices may splurge on new luxury cars in anticipation of lucrative professional football contracts. Or workers who become fearful of losing their jobs may reduce their demand for, say, vacation travel.

In summary, an *increase* in demand—the decision by consumers to buy larger quantities of a product at each possible price—may be caused by:

- A favorable change in consumer tastes.
- An increase in the number of buyers.
- Rising incomes if the product is a normal good.
- Falling incomes if the product is an inferior good.
- An increase in the price of a substitute good.
- A decrease in the price of a complementary good.
- A new consumer expectation that either prices or income will be higher in the future.

You should "reverse" these generalizations to explain a *decrease* in demand. Table 3.1 provides additional illustrations of the determinants of demand.

Changes in Quantity Demanded

A *change in demand* must not be confused with a *change in quantity demanded.* A **change in demand** is a shift of the demand curve to the right (an increase in demand) or to the left (a decrease in demand). It occurs because the consumer's state of mind about purchasing the product has been altered in response to a change in one or more of the determinants of demand. Recall that "demand" is a schedule or a curve; therefore, a "change in demand" means a change in the schedule and a shift of the curve.

In contrast, a **change in quantity demanded** is a movement from one point to another point—from one price-quantity combination to another—on a fixed demand curve. The cause of such a change is an increase or

TABLE 3.1 **Determinants of Demand: Factors That Shift the Demand Curve**

Determinant	Examples
Change in buyer tastes	Physical fitness rises in popularity, increasing the demand for jogging shoes and bicycles; cell phone popularity rises, reducing the demand for landline phones.
Change in number of buyers	A decline in the birthrate reduces the demand for children's toys.
Change in income	A rise in incomes increases the demand for normal goods such as restaurant meals, sports tickets, and necklaces while reducing the demand for inferior goods such as cabbage, turnips, and inexpensive wine.
Change in the prices of related goods	A reduction in airfares reduces the demand for bus transportation (substitute goods); a decline in the price of DVD players increases the demand for DVD movies (complementary goods).
Change in consumer expectations	Inclement weather in South America creates an expectation of higher future coffee bean prices, thereby increasing today's demand for coffee beans.

decrease in the price of the product under consideration. In the table in Figure 3.3, for example, a decline in the price of corn from $5 to $4 will increase the quantity demanded of corn from 2,000 to 4,000 bushels.

In Figure 3.3 the shift of the demand curve D_1 to either D_2 or D_3 is a change in demand. But the movement from point *a* to point *b* on curve D_1 represents a change in quantity demanded*:* Demand has not changed; it is the entire curve, and it remains fixed in place.

QUICK REVIEW 3.1

- Demand is a schedule or a curve showing the amount of a product that buyers are willing and able to purchase, in a particular time period, at each possible price in a series of prices.
- The law of demand states that, other things equal, the quantity of a good purchased varies inversely with its price.
- The demand curve shifts because of changes in (a) consumer tastes, (b) the number of buyers in the market, (c) consumer income, (d) the prices of substitute or complementary goods, and (e) consumer expectations.
- A change in demand is a shift of the demand curve; a change in quantity demanded is a movement from one point to another on a fixed demand curve.

Supply

LO3.3 Describe *supply* and explain how it can change.

Supply is a schedule or curve showing the various amounts of a product that producers are willing and able to make available for sale at each of a series of possible prices during a specific period.[2] The table in Figure 3.4 is a hypothetical **supply schedule** for a single producer of corn. It shows the quantities of corn that will be supplied at various prices, other things equal.

Law of Supply

The table in Figure 3.4 shows that a positive or direct relationship prevails between price and quantity supplied. As price rises, the quantity supplied rises; as price falls, the quantity supplied falls. This relationship is called the **law of supply.** A supply schedule tells us that, other things equal, firms will produce and offer for sale more of their product at a high price than at a low price. This, again, is basically common sense.

Price is an obstacle from the standpoint of the consumer, who is on the paying end. The higher the price, the less the consumer will buy. But the supplier is on the receiving end of the product's price. To a supplier, price represents *revenue*, which serves as an incentive to produce and sell a product. The higher the price, the greater this incentive and the greater the quantity supplied.

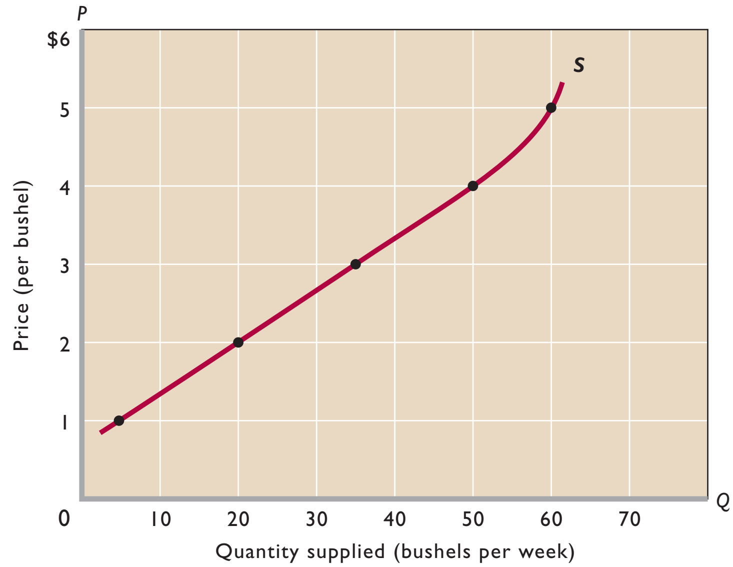
Consider a farmer who is deciding on how much corn to plant. As corn prices rise, as shown in the table in Figure 3.4, the farmer finds it profitable to plant more corn. And the higher corn prices enable the farmer to cover the increased costs associated with more intensive cultivation and the use of more seed, fertilizer, and pesticides. The overall result is more corn.

Now consider a manufacturer. Beyond some quantity of production, manufacturers usually encounter increases in *marginal cost*—the added cost of producing one more unit of output. Certain productive resources—in particular, the firm's plant and machinery—cannot be expanded quickly, so the firm uses more of other resources such as labor to produce more output. But as labor becomes more abundant relative to the fixed plant and equipment, the additional workers have relatively less space and access to equipment. For example, the added workers may have to wait to gain access to machines. As a result, each added worker produces less added output, and the marginal cost of successive units of output rises accordingly. The firm will not produce the

[2]This definition is worded to apply to product markets. To adjust it to apply to resource markets, substitute "resource" for "product" and "owners" for "producers."

FIGURE 3.4 An individual producer's supply of corn. Because price and quantity supplied are directly related, the supply curve for an individual producer graphs as an upsloping curve. Other things equal, producers will offer more of a product for sale as its price rises and less of the product for sale as its price falls.

Supply of Corn	
Price per Bushel	Quantity Supplied per Week
$5	60
4	50
3	35
2	20
1	5



more costly units unless it receives a higher price for them. Again, price and quantity supplied are directly related.

The Supply Curve

As with demand, it is convenient to represent individual supply graphically. In Figure 3.4, curve S is the **supply curve** that corresponds with the price–quantity supplied data in the accompanying table. The upward slope of the curve reflects the law of supply—producers offer more of a good, service, or resource for sale as its price rises. The relationship between price and quantity supplied is positive, or direct.

Market Supply

Market supply is derived from individual supply in exactly the same way that market demand is derived from individual demand. We sum the quantities supplied by each producer at each price. That is, we obtain the market supply curve by "horizontally adding" the supply curves of the individual producers. The price–quantity supplied data in the table accompanying Figure 3.5 are for an assumed 200 identical producers in the market, each willing to supply corn according to the supply schedule shown in Figure 3.4. Curve S_1 in Figure 3.5 is a graph of the market supply data. Note that the values of the axes in Figure 3.5 are the same as those used in our graph of market demand (Figure 3.3). The only difference is that we change the label on the horizontal axis from "quantity demanded" to "quantity supplied."

Determinants of Supply

In constructing a supply curve, we assume that price is the most significant influence on the quantity supplied of any product. But other factors (the "other things equal") can and do affect supply. The supply curve is drawn on the assumption that these other things are fixed and do not change. If one of them does change, a *change in supply* will occur, meaning that the entire supply curve will shift.

The basic **determinants of supply** are (1) resource prices, (2) technology, (3) taxes and subsidies, (4) prices of other goods, (5) producer expectations, and (6) the number of sellers in the market. A change in any one or more of these determinants of supply, or *supply shifters*, will move the supply curve for a product either right or left. A shift to the *right*, as from S_1 to S_2 in Figure 3.5, signifies an *increase* in supply: Producers supply larger quantities of the product at each possible price. A shift to the *left*, as from S_1 to S_3, indicates a *decrease* in supply: Producers offer less output at each price.

Changes in Supply

Let's consider how changes in each of the determinants affect supply. The key idea is that costs are a major factor underlying supply curves; anything that affects costs (other than changes in output itself) usually shifts the supply curve.

Resource Prices The prices of the resources used in the production process help determine the costs of production incurred by firms. Higher *resource* prices raise

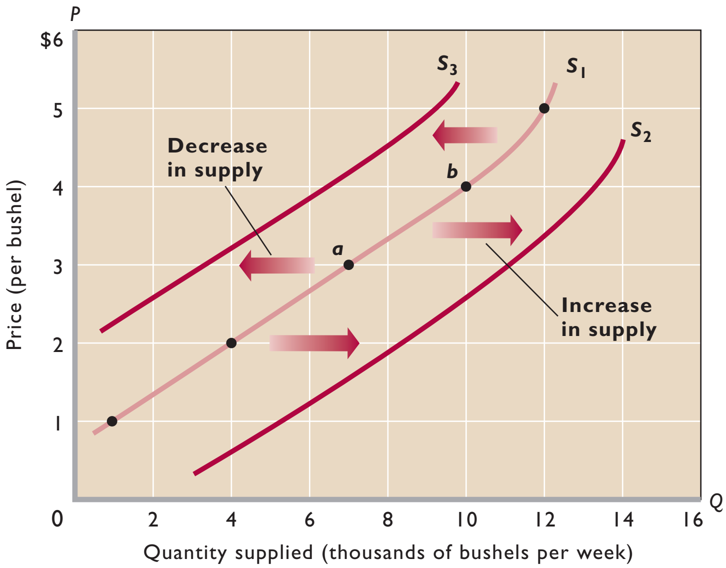


FIGURE 3.5 Changes in the supply of corn. A change in one or more of the determinants of supply causes a change in supply. An increase in supply is shown as a rightward shift of the supply curve, as from S_1 to S_2. A decrease in supply is depicted as a leftward shift of the curve, as from S_1 to S_3. In contrast, a change in the *quantity supplied* is caused by a change in the product's price and is shown by a movement from one point to another, as from *b* to *a* on fixed supply curve S_1.

Market Supply of Corn, 200 Producers, (S_1)	
(1) Price per Bushel	(2) Total Quantity Supplied per Week
$5	12,000
4	10,000
3	7,000
2	4,000
1	1,000

production costs and, assuming a particular *product* price, squeeze profits. That reduction in profits reduces the incentive for firms to supply output at each product price. For example, an increase in the price of sand, crushed rock, or Portland cement will increase the cost of producing concrete and reduce its supply.

In contrast, lower *resource* prices reduce production costs and increase profits. So when resource prices fall, firms supply greater output at each product price. For example, a decrease in the price of iron ore will decrease the price of steel.

Technology Improvements in technology (techniques of production) enable firms to produce units of output with fewer resources. Because resources are costly, using fewer of them lowers production costs and increases supply. Example: Technological advances in producing flat-panel computer monitors have greatly reduced their cost. Thus, manufacturers will now offer more such monitors than previously at the various prices; the supply of flat-panel monitors has increased.

Taxes and Subsidies Businesses treat most taxes as costs. An increase in sales or property taxes will increase production costs and reduce supply. In contrast, subsidies are "taxes in reverse." If the government subsidizes the production of a good, it in effect lowers the producers' costs and increases supply.

Prices of Other Goods Firms that produce a particular product, say, soccer balls, can sometimes use their plant and equipment to produce alternative goods, say, basketballs and volleyballs. The higher prices of these "other goods" may entice soccer ball producers to switch production to those other goods in order to increase profits. This *substitution in production* results in a decline in the supply of soccer balls. Alternatively, when the prices of basketballs and volleyballs decline relative to the price of soccer balls, producers of those goods may decide to produce more soccer balls instead, increasing their supply.

Producer Expectations Changes in expectations about the future price of a product may affect the producer's current willingness to supply that product. It is difficult, however, to generalize about how a new expectation of higher prices affects the present supply of a product. Farmers anticipating a higher wheat price in the future might withhold some of their current wheat harvest from the market, thereby causing a decrease in the current supply of wheat. In contrast, in many types of manufacturing industries, newly formed expectations that price will increase may induce firms to add another shift of workers or to expand their production facilities, causing current supply to increase.

Number of Sellers Other things equal, the larger the number of suppliers, the greater the market supply. As more firms enter an industry, the supply curve shifts to the right. Conversely, the smaller the number of firms in the industry, the less the market supply. This means that as firms leave an industry, the supply curve shifts to the left. Example: The United States and Canada have imposed restrictions on haddock fishing to replenish dwindling stocks. As part of that policy, the federal government has bought the boats of some of the haddock fishers as a way of putting

TABLE 3.2 Determinants of Supply: Factors That Shift the Supply Curve

Determinant	Examples
Change in resource prices	A decrease in the price of microchips increases the supply of computers; an increase in the price of crude oil reduces the supply of gasoline.
Change in technology	The development of more effective wireless technology increases the supply of cell phones.
Changes in taxes and subsidies	An increase in the excise tax on cigarettes reduces the supply of cigarettes; a decline in subsidies to state universities reduces the supply of higher education.
Change in prices of other goods	An increase in the price of cucumbers decreases the supply of watermelons.
Change in producer expectations	An expectation of a substantial rise in future log prices decreases the supply of logs today.
Change in number of suppliers	An increase in the number of tattoo parlors increases the supply of tattoos; the formation of women's professional basketball leagues increases the supply of women's professional basketball games.

them out of business and decreasing the catch. The result has been a decline in the market supply of haddock.

Table 3.2 is a checklist of the determinants of supply, along with further illustrations.

Changes in Quantity Supplied

The distinction between a *change in supply* and a *change in quantity supplied* parallels the distinction between a change in demand and a change in quantity demanded. Because supply is a schedule or curve, a **change in supply** means a change in the schedule and a shift of the curve. An increase in supply shifts the curve to the right; a decrease in supply shifts it to the left. The cause of a change in supply is a change in one or more of the determinants of supply.

In contrast, a **change in quantity supplied** is a movement from one point to another on a fixed supply curve. The cause of such a movement is a change in the price of the specific product being considered.

Consider supply curve S_1 in Figure 3.5. A decline in the price of corn from \$4 to \$3 decreases the quantity of corn supplied per week from 10,000 to 7,000 bushels. This movement from point *b* to point *a* along S_1 is a change in quantity supplied, not a change in supply. Supply is the full schedule of prices and quantities shown, and this schedule does not change when the price of corn changes.

QUICK REVIEW 3.2

- A supply schedule or curve shows that, other things equal, the quantity of a good supplied varies directly with its price.
- The supply curve shifts because of changes in (a) resource prices, (b) technology, (c) taxes or subsidies, (d) prices of other goods, (e) expectations of future prices, and (f) the number of suppliers.
- A change in supply is a shift of the supply curve; a change in quantity supplied is a movement from one point to another on a fixed supply curve.

Market Equilibrium

LO3.4 Relate how supply and demand interact to determine market equilibrium.

With our understanding of demand and supply, we can now show how the decisions of buyers of corn and sellers of corn interact to determine the equilibrium price and quantity of corn. In the table in Figure 3.6, columns 1 and 2 repeat the market supply of corn (from the table in Figure 3.5), and columns 2 and 3 repeat the market demand for corn (from the table in Figure 3.3). We assume this is a competitive market so that neither buyers nor sellers can set the price.

Equilibrium Price and Quantity

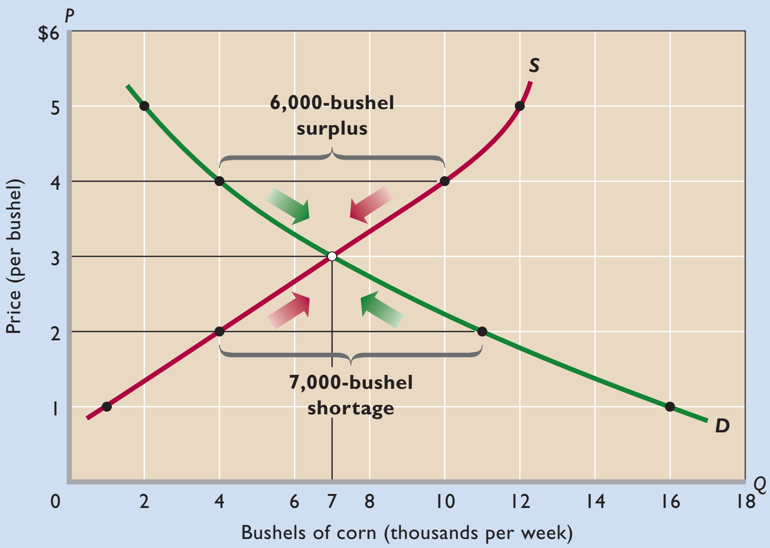
We are looking for the equilibrium price and equilibrium quantity. The **equilibrium price** (or *market-clearing price*) is the price where the intentions of buyers and sellers match. It is the price where quantity demanded equals quantity supplied. The table in Figure 3.6 reveals that at \$3, *and only at that price*, the number of bushels of corn that sellers wish to sell (7,000) is identical to the number consumers want to buy (also 7,000). At \$3 and 7,000 bushels of corn, there is neither a shortage nor a surplus of corn. So 7,000 bushels of corn is the **equilibrium quantity:** the quantity at which the intentions of buyers and sellers match, so that the quantity demanded and the quantity supplied are equal.

Graphically, the equilibrium price is indicated by the intersection of the supply curve and the demand curve in **Figure 3.6 (Key Graph).** (The horizontal axis now measures both quantity demanded and quantity supplied.) With neither a shortage nor a surplus at \$3, the market is *in equilibrium*, meaning "in balance" or "at rest."

Competition among buyers and among sellers drives the price to the equilibrium price; once there, it will remain there unless it is subsequently disturbed by changes in demand or supply (shifts of the curves). To better understand the uniqueness of the equilibrium price, let's consider other

KEY GRAPH

FIGURE 3.6 Equilibrium price and quantity. The intersection of the downsloping demand curve *D* and the upsloping supply curve *S* indicates the equilibrium price and quantity, here \$3 and 7,000 bushels of corn. The shortages of corn at below-equilibrium prices (for example, 7,000 bushels at \$2) drive up price. The higher prices increase the quantity supplied and reduce the quantity demanded until equilibrium is achieved. The surpluses caused by above-equilibrium prices (for example, 6,000 bushels at \$4) push price down. As price drops, the quantity demanded rises and the quantity supplied falls until equilibrium is established. At the equilibrium price and quantity, there are neither shortages nor surpluses of corn.



Market Supply of and Demand for Corn

(1) Total Quantity Supplied per Week	(2) Price per Bushel	(3) Total Quantity Demanded per Week	(4) Surplus (+) or Shortage (−)*
12,000	\$5	2,000	+10,000 ↓
10,000	4	4,000	+6,000 ↓
7,000	3	7,000	0
4,000	2	11,000	−7,000 ↑
1,000	1	16,000	−15,000 ↑

*Arrows indicate the effect on price.

QUICK QUIZ FOR FIGURE 3.6

1. Demand curve *D* is downsloping because:

a. producers offer less of a product for sale as the price of the product falls.

b. lower prices of a product create income and substitution effects that lead consumers to purchase more of it.

c. the larger the number of buyers in a market, the lower the product price.

d. price and quantity demanded are directly (positively) related.

2. Supply curve *S*:

a. reflects an inverse (negative) relationship between price and quantity supplied.

b. reflects a direct (positive) relationship between price and quantity supplied.

c. depicts the collective behavior of buyers in this market.

d. shows that producers will offer more of a product for sale at a low product price than at a high product price.

3. At the \$3 price:

a. quantity supplied exceeds quantity demanded.

b. quantity demanded exceeds quantity supplied.

c. the product is abundant and a surplus exists.

d. there is no pressure on price to rise or fall.

4. At price \$5 in this market:

a. there will be a shortage of 10,000 units.

b. there will be a surplus of 10,000 units.

c. quantity demanded will be 12,000 units.

d. quantity demanded will equal quantity supplied.

Answers: 1. b; 2. b; 3. d; 4. b

prices. At any above-equilibrium price, quantity supplied exceeds quantity demanded. For example, at the \$4 price, sellers will offer 10,000 bushels of corn, but buyers will purchase only 4,000. The \$4 price encourages sellers to offer lots of corn but discourages many consumers from buying it. The result is a **surplus** (or *excess supply*) of 6,000 bushels. If corn sellers produced them all, they would find themselves with 6,000 unsold bushels of corn.

Surpluses drive prices down. Even if the \$4 price existed temporarily, it could not persist. The large surplus would prompt competing sellers to lower the price to encourage buyers to take the surplus off their hands. As the price fell, the incentive to produce corn would decline and the incentive for consumers to buy corn would increase. As shown in Figure 3.6, the market would move to its equilibrium at \$3.

Any price below the \$3 equilibrium price would create a shortage; quantity demanded would exceed quantity supplied. Consider a \$2 price, for example. We see both from column 2 of the table and from the demand curve in Figure 3.6 that quantity demanded exceeds quantity supplied at that price. The result is a **shortage** (or *excess demand*) of 7,000 bushels of corn. The $2 price discourages sellers from devoting resources to corn and encourages consumers to desire more bushels than are available. The $2 price cannot persist as the equilibrium price. Many consumers who want to buy corn at this price will not obtain it. They will express a willingness to pay more than $2 to get corn. Competition among these buyers will drive up the price, eventually to the $3 equilibrium level. Unless disrupted by changes of supply or demand, this $3 price of corn will continue to prevail.

CONSIDER THIS . . .

Ticket Scalping: A Bum Rap!

Ticket prices for athletic events and musical concerts are usually set far in advance of the events. Sometimes the original ticket price is too low to be the equilibrium price. Lines form at the ticket window and a severe shortage of tickets occurs at the printed price. What happens next? Buyers who are willing to pay more than the original price bid up the ticket price in resale ticket markets.

Tickets sometimes get resold for much greater amounts than the original price—market transactions known as "scalping." For example, an original buyer may resell a $75 ticket to a concert for $200. Reporters sometimes denounce scalpers for "ripping off" buyers by charging "exorbitant" prices.

But is scalping really a rip-off? We must first recognize that such ticket resales are voluntary transactions. If both buyer and seller did not expect to gain from the exchange, it would not occur! The seller must value the $200 more than seeing the event, and the buyer must value seeing the event at $200 or more. So there are no losers or victims here: Both buyer and seller benefit from the transaction. The scalping market simply redistributes assets (game or concert tickets) from those who would rather have the money (and the other things that the money can buy) to those who would rather have the tickets.

Does scalping impose losses or injury on the sponsors of the event? If the sponsors are injured, it is because they initially priced tickets below the equilibrium level. Perhaps they did this to create a long waiting line and the attendant news media publicity. Alternatively, they may have had a genuine desire to keep tickets affordable for lower-income, ardent fans. In either case, the event sponsors suffer an opportunity cost in the form of less ticket revenue than they might have otherwise received. But such losses are self-inflicted and separate and distinct from the fact that some tickets are later resold at a higher price.

So is ticket scalping undesirable? Not on economic grounds! It is an entirely voluntary activity that benefits both sellers and buyers.

Rationing Function of Prices

The ability of the competitive forces of supply and demand to establish a price at which selling and buying decisions are consistent is called the rationing function of prices. In our case, the equilibrium price of $3 clears the market, leaving no burdensome surplus for sellers and no inconvenient shortage for potential buyers. And it is the combination of freely made individual decisions that sets this market-clearing price. In effect, the market outcome says that all buyers who are willing and able to pay $3 for a bushel of corn will obtain it; all buyers who cannot or will not pay $3 will go without corn. Similarly, all producers who are willing and able to offer corn for sale at $3 a bushel will sell it; all producers who cannot or will not sell for $3 per bushel will not sell their product.

Efficient Allocation

A competitive market such as that we have described not only rations goods to consumers but also allocates society's resources efficiently to the particular product. Competition among corn producers forces them to use the best technology and right mix of productive resources. If they didn't, their costs would be too high relative to the market price, and they would be unprofitable. The result is **productive efficiency:** the production of any particular good in the least costly way. When society produces corn at the lowest achievable per-unit cost, it is expending the least-valued combination of resources to produce that product and therefore is making available more-valued resources to produce other desired goods. Suppose society has only $100 worth of resources available. If it can produce a bushel of corn using $3 of those resources, then it will have available $97 of resources remaining to produce other goods. This is clearly better than producing the corn for $5 and having only $95 of resources available for the alternative uses.

Competitive markets also produce **allocative efficiency:** the *particular mix* of goods and services most highly valued by society (minimum-cost production assumed). For example, society wants land suitable for growing corn used for that purpose, not to grow dandelions. It wants diamonds to be used for jewelry, not crushed up and used as an additive to give concrete more sparkle. It wants iPods and MP4 players, not cassette players and tapes. Moreover, society does not want to devote all its resources to corn, diamonds, and portable digital media players. It wants to assign some resources to wheat, gasoline, and cell phones. Competitive markets make those allocatively efficient assignments.

The equilibrium price and quantity in competitive markets usually produce an assignment of resources that is "right" from an economic perspective. Demand essentially reflects the marginal benefit (MB) of the good, based on the utility received. Supply reflects the marginal cost (MC) of producing the good. The market ensures that firms produce all units of goods for which MB exceeds MC and no units for which MC exceeds MB. At the intersection of the demand and supply curves, MB equals MC and allocative efficiency results. As economists say, there is neither an "underallocaton of resources" nor an "overallocation of resources" to the product.

Changes in Supply, Demand, and Equilibrium

LO3.5 Explain how changes in supply and demand affect equilibrium prices and quantities.

We know that demand might change because of fluctuations in consumer tastes or incomes, changes in consumer expectations, or variations in the prices of related goods. Supply might change in response to changes in resource prices, technology, or taxes. What effects will such changes in supply and demand have on equilibrium price and quantity?

Changes in Demand

Suppose that the supply of some good (for example, health care) is constant and demand increases, as shown in Figure 3.7a. As a result, the new intersection of the supply and demand curves is at higher values on both the price and the quantity axes. Clearly, an increase in demand raises both equilibrium price and equilibrium quantity. Conversely, a decrease in demand such as that shown in Figure 3.7b reduces both equilibrium price and equilibrium quantity. (The value of graphical analysis is now apparent: We need not fumble with columns of figures to determine the outcomes; we need only compare the new and the old points of intersection on the graph.)

Changes in Supply

What happens if the demand for some good (for example, flash drives) is constant but supply increases, as in Figure 3.7c? The new intersection of supply and demand is located at a lower equilibrium price but at a higher equilibrium quantity. An increase in supply reduces equilibrium price but increases equilibrium quantity. In contrast, if supply decreases, as in Figure 3.7d, equilibrium price rises while equilibrium quantity declines.

Complex Cases

When both supply and demand change, the effect is a combination of the individual effects.

Supply Increase; Demand Decrease What effect will a supply increase and a demand decrease for some good (for example, apples) have on equilibrium price? Both changes decrease price, so the net result is a price drop greater than that resulting from either change alone.

What about equilibrium quantity? Here the effects of the changes in supply and demand are opposed: the increase

CONSIDER THIS . . .

Salsa and Coffee Beans

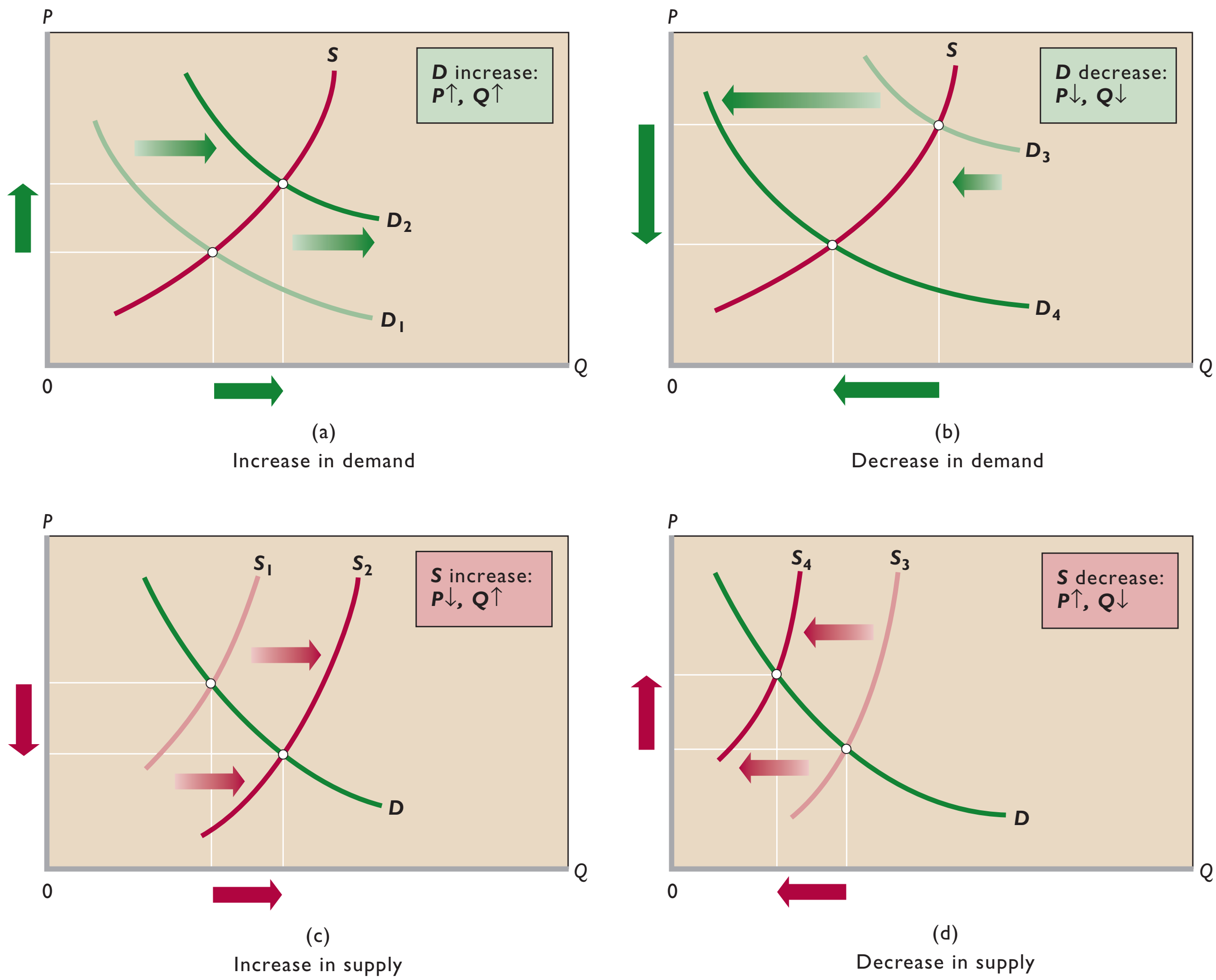
If you forget the other-things-equal assumption, you can encounter situations that *seem* to be in conflict with the laws of demand and supply. For example, suppose salsa manufacturers sell 1 million bottles of salsa at $4 a bottle in one year; 2 million bottles at $5 in the next year; and 3 million at $6 in the year thereafter. Price and quantity purchased vary directly, and these data seem to be at odds with the law of demand.

But there is no conflict here; the data do not refute the law of demand. The catch is that the law of demand's other-things-equal assumption has been violated over the three years in the example. Specifically, because of changing tastes and rising incomes, the demand for salsa has increased sharply, as in Figure 3.7a. The result is higher prices *and* larger quantities purchased.

Another example: The price of coffee beans occasionally shoots upward at the same time that the quantity of coffee beans harvested declines. These events seemingly contradict the direct relationship between price and quantity denoted by supply. The catch again is that the other-things-equal assumption underlying the upsloping supply curve is violated. Poor coffee harvests decrease supply, as in Figure 3.7d, increasing the equilibrium price of coffee and reducing the equilibrium quantity.

The laws of demand and supply are not refuted by observations of price and quantity made over periods of time in which either demand or supply curves shift.

FIGURE 3.7 Changes in demand and supply and the effects on price and quantity. The increase in demand from D_1 to D_2 in (a) increases both equilibrium price and equilibrium quantity. The decrease in demand from D_3 to D_4 in (b) decreases both equilibrium price and equilibrium quantity. The increase in supply from S_1 to S_2 in (c) decreases equilibrium price and increases equilibrium quantity. The decline in supply from S_3 to S_4 in (d) increases equilibrium price and decreases equilibrium quantity. The boxes in the top right corners summarize the respective changes and outcomes. The upward arrows in the boxes signify increases in equilibrium price (*P*) and equilibrium quantity (*Q*); the downward arrows signify decreases in these items.



in supply increases equilibrium quantity, but the decrease in demand reduces it. The direction of the change in equilibrium quantity depends on the relative sizes of the changes in supply and demand. If the increase in supply is larger than the decrease in demand, the equilibrium quantity will increase. But if the decrease in demand is greater than the increase in supply, the equilibrium quantity will decrease.

Supply Decrease; Demand Increase A decrease in supply and an increase in demand for some good (for example, gasoline) both increase price. Their combined effect is an increase in equilibrium price greater than that caused by either change separately. But their effect on the equilibrium quantity is again indeterminate, depending on the relative sizes of the changes in supply and demand. If the decrease in supply is larger than the increase in demand, the equilibrium quantity will decrease. In contrast, if the increase in demand is greater than the decrease in supply, the equilibrium quantity will increase.

Supply Increase; Demand Increase What if supply and demand both increase for some good (for example, cell phones)? A supply increase drops equilibrium price, while a demand increase boosts it. If the increase in supply is greater than the increase in demand, the equilibrium price will fall. If the opposite holds, the equilibrium price will rise.

The effect on equilibrium quantity is certain: The increases in supply and demand both raise the equilibrium

TABLE 3.3 Effects of Changes in Both Supply and Demand

Change in Supply	Change in Demand	Effect on Equilibrium Price	Effect on Equilibrium Quantity
1. Increase	Decrease	Decrease	Indeterminate
2. Decrease	Increase	Increase	Indeterminate
3. Increase	Increase	Indeterminate	Increase
4. Decrease	Decrease	Indeterminate	Decrease

quantity. Therefore, the equilibrium quantity will increase by an amount greater than that caused by either change alone.

Supply Decrease; Demand Decrease What about decreases in both supply and demand for some good (for example, new homes)? If the decrease in supply is greater than the decrease in demand, equilibrium price will rise. If the reverse is true, equilibrium price will fall. Because the decreases in supply and demand each reduce equilibrium quantity, we can be sure that equilibrium quantity will fall.

Table 3.3 summarizes these four cases. To understand them fully, you should draw supply and demand diagrams for each case to confirm the effects listed in this table.

Special cases arise when a decrease in demand and a decrease in supply, or an increase in demand and an increase in supply, exactly cancel out. In both cases, the net effect on equilibrium price will be zero; price will not change.

The optional appendix accompanying this chapter provides additional examples of situations in which both supply and demand change at the same time.

Application: Government-Set Prices

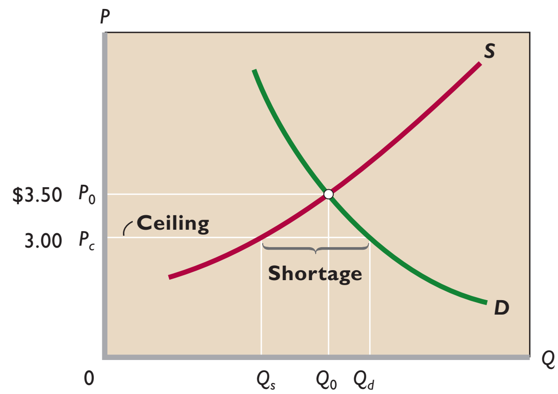
LO3.6 Identify what government-set prices are and how they can cause product surpluses and shortages.

Prices in most markets are free to rise or fall to their equilibrium levels, no matter how high or low those levels might be. However, government sometimes concludes that supply and demand will produce prices that are unfairly high for buyers or unfairly low for sellers. So government may place legal limits on how high or low a price or prices may go. Is that a good idea?

Price Ceilings on Gasoline

A **price ceiling** sets the maximum legal price a seller may charge for a product or service. A price at or below the ceiling is legal; a price above it is not. The rationale for establishing price ceilings (or ceiling prices) on specific products is that they purportedly enable consumers to obtain some "essential" good or service that they could not afford at the equilibrium price. Examples are rent controls and usury laws, which specify maximum "prices" in the forms of rent and interest that can be charged to borrowers.

FIGURE 3.8 **A price ceiling.** A price ceiling is a maximum legal price such as P_c. When the ceiling price is below the equilibrium price, a persistent product shortage results. Here that shortage is shown by the horizontal distance between Q_d and Q_s.



Graphical Analysis We can easily show the effects of price ceilings graphically. Suppose that rapidly rising world income boosts the purchase of automobiles and shifts the demand for gasoline to the right so that the market equilibrium price reaches $3.50 per gallon, shown as P_0 in Figure 3.8. The rapidly rising price of gasoline greatly burdens low- and moderate-income households, which pressure government to "do something." To keep gasoline prices down, the government imposes a ceiling price P_c of $3 per gallon. To impact the market, a price ceiling must be below the equilibrium price. A ceiling price of $4, for example, would have had no effect on the price of gasoline in the current situation.

What are the effects of this $3 ceiling price? The rationing ability of the free market is rendered ineffective. Because the ceiling price P_c is below the market-clearing price P_0, there is a lasting shortage of gasoline. The quantity of gasoline demanded at P_c is Q_d and the quantity supplied is only Q_s; a persistent excess demand or shortage of amount $Q_d - Q_s$ occurs.

The price ceiling P_c prevents the usual market adjustment in which competition among buyers bids up price, inducing more production and rationing some buyers out of the market. That process would normally continue until the shortage disappeared at the equilibrium price and quantity, P_0 and Q_0.

By preventing these market adjustments from occurring, the price ceiling poses two related problems.

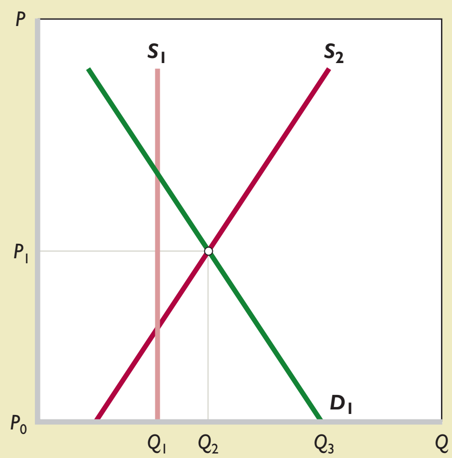
Rationing Problem How will the available supply Q_s be apportioned among buyers who want the greater

LAST WORD

A Legal Market for Human Organs?

A Legal Market Might Eliminate the Present Shortage of Human Organs for Transplant. But There Are Many Serious Objections to "Turning Human Body Parts into Commodities" for Purchase and Sale.

It has become increasingly commonplace in medicine to transplant kidneys, lungs, livers, corneas, pancreases, and hearts from deceased individuals to those whose organs have failed or are failing. But surgeons and many of their patients face a growing problem: There are shortages of donated organs available for transplant. Not everyone who needs a transplant can get one. In 2012, there were 116,000 Americans on the waiting list for transplants. Indeed, an inadequate supply of donated organs causes an estimated 6,900 deaths in the United States each year.



Why Shortages? Seldom do we hear of shortages of desired goods in market economies. What is different about organs for transplant? One difference is that no legal market exists for human organs. To understand this situation, observe the demand curve D_1 and supply curve S_1 in the accompanying figure. The downward slope of the demand curve tells us that if there were a market for human organs, the quantity of organs demanded would be greater at lower prices than at higher prices. Vertical supply curve S_1 represents the fixed quantity of human organs now donated via consent before death. Because the price of these donated organs is in effect zero, quantity demanded Q_3 exceeds quantity supplied Q_1. The shortage of $Q_3 - Q_1$ is rationed through a waiting list of those in medical need of transplants. Many people die while still on the waiting list.

Use of a Market A market for human organs would increase the incentive to donate organs. Such a market might work like this: An individual might specify in a legal document that he or she is willing to sell one or more usable human organs upon death or near-death. The person could specify where the money from the sale would go, for example, to family, a church, an educational institution, or a charity. Firms would then emerge to purchase

amount Q_d? Should gasoline be distributed on a first-come, first-served basis, that is, to those willing and able to get in line the soonest or stay in line the longest? Or should gas stations distribute it on the basis of favoritism? Since an unregulated shortage does not lead to an equitable distribution of gasoline, the government must establish some formal system for rationing it to consumers. One option is to issue ration coupons, which authorize bearers to purchase a fixed amount of gasoline per month. The rationing system might entail first the printing of coupons for Q_s gallons of gasoline and then the equal distribution of the coupons among consumers so that the wealthy family of four and the poor family of four both receive the same number of coupons.

Black Markets But ration coupons would not prevent a second problem from arising. The demand curve in Figure 3.8 reveals that many buyers are willing to pay more than the ceiling price P_c. And, of course, it is more profitable for gasoline stations to sell at prices above the ceiling. Thus, despite a sizable enforcement bureaucracy that would have to accompany the price controls, *black markets* in which gasoline is illegally bought and sold at prices above the legal limits will flourish. Counterfeiting of ration coupons will also be a problem. And since the price of gasoline is now "set by government," government might face political pressure to set the price even lower.

Rent Controls

About 200 cities in the United States, including New York City, Boston, and San Francisco, have at one time or another enacted rent controls: maximum rents established by law (or, more recently, maximum rent increases for existing tenants). Such laws are well intended. Their goals are to protect low-income families from escalating rents caused by perceived housing shortages and to make housing more affordable to the poor.

organs and resell them where needed for profit. Under such a system, the supply curve of usable organs would take on the normal upward slope of typical supply curves. The higher the expected price of an organ, the greater the number of people who would be willing to have their organs sold at death. Suppose that the supply curve is S_2 in the figure. At the equilibrium price P_1, the number of organs made available for transplant (Q_2) would equal the number purchased for transplant (also Q_2). In this generalized case, the shortage of organs would be eliminated and, of particular importance, the number of organs available for transplanting would rise from Q_1 to Q_2. This means more lives would be saved and enhanced than under the present donor system.

Objections In view of this positive outcome, why is there no such market for human organs? Critics of market-based solutions have two main objections. The first is a moral objection: Critics feel that turning human organs into commodities commercializes human beings and diminishes the special nature of human life. They say there is something unseemly about selling and buying body organs as if they were bushels of wheat or ounces of gold. (There is, however, a market for blood!) Moreover, critics note that the market would ration the available organs (as represented by Q_2 in the figure) to people who either can afford them (at P_1) or have health insurance for transplants. The poor and uninsured would be left out.

Second, a health-cost objection suggests that a market for body organs would greatly increase the cost of health care. Rather than obtaining freely donated (although "too few") body organs, patients or their insurance companies would have to pay market prices for them, further increasing the cost of medical care.

Rebuttal Supporters of market-based solutions to organ shortages point out that the laws against selling organs are simply driving the market underground. Worldwide, an estimated $1 billion-per-year illegal market in human organs has emerged. As in other illegal markets, the unscrupulous tend to thrive. This fact is dramatized by the accompanying photo, in which four Pakistani villagers show off their scars after they each sold a kidney to pay off debts. Supporters say that legalization of the market for human organs would increase organ supply from legal sources, drive down the price of organs, and reduce the abuses such as those now taking place in illegal markets.

What have been the actual economic effects? On the demand side, the below-equilibrium rents attract a larger number of renters. Some are locals seeking to move into their own places after sharing housing with friends or family. Others are outsiders attracted into the area by the artificially lower rents. But a large problem occurs on the supply side. Price controls make it less attractive for landlords to offer housing on the rental market. In the short run, owners may sell their rental units or convert them to condominiums. In the long run, low rents make it unprofitable for owners to repair or renovate their rental units. (Rent controls are one cause of the many abandoned apartment buildings found in larger cities.) Also, insurance companies, pension funds, and other potential new investors in housing will find it more profitable to invest in office buildings, shopping malls, or motels, where rents are not controlled.

In brief, rent controls distort market signals and thus resources are misallocated: Too few resources are allocated to rental housing and too many to alternative uses. Ironically, although rent controls are often legislated to lessen the effects of perceived housing shortages, controls in fact are a primary cause of such shortages. For that reason, most American cities either have abandoned or are in the process of dismantling rent controls.

Price Floors on Wheat

A **price floor** is a minimum price fixed by the government. A price at or above the price floor is legal; a price below it is not. Price floors above equilibrium prices are usually invoked when society feels that the free functioning of the market system has not provided a sufficient income for certain groups of resource suppliers or producers. Supported prices for agricultural products and current minimum wages are two examples of price (or wage) floors. Let's look at the former.

Suppose that many farmers have extremely low incomes when the price of wheat is at its equilibrium value of $2 per bushel. The government decides to help out by establishing a legal price floor or price support of $3 per bushel.

What will be the effects? At any price above the equilibrium price, quantity supplied will exceed quantity demanded—that is, there will be a persistent excess supply or surplus of the product. Farmers will be willing to produce and offer for sale more than private buyers are willing to purchase at the price floor. As we saw with a price ceiling, an imposed legal price disrupts the rationing ability of the free market.

Graphical Analysis Figure 3.9 illustrates the effect of a price floor graphically. Suppose that *S* and *D* are the supply and demand curves for wheat. Equilibrium price and quantity are P_0 and Q_0, respectively. If the government imposes a price floor of P_f, farmers will produce Q_s but private buyers will purchase only Q_d. The surplus is the excess of Q_s over Q_d.

The government may cope with the surplus resulting from a price floor in two ways:

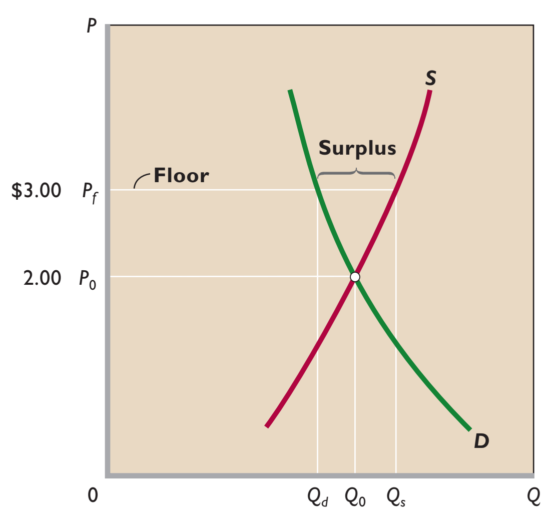
- It can restrict supply (for example, by instituting acreage allotments by which farmers agree to take a certain amount of land out of production) or increase demand (for example, by researching new uses for the product involved). These actions may reduce the difference between the equilibrium price and the price floor and that way reduce the size of the resulting surplus.
- If these efforts are not wholly successful, then the government must purchase the surplus output at the $3 price (thereby subsidizing farmers) and store or otherwise dispose of it.

Additional Consequences Price floors such as P_f in Figure 3.9 not only disrupt the rationing ability of prices but distort resource allocation. Without the price floor, the $2 equilibrium price of wheat would cause financial losses and force high-cost wheat producers to plant other crops or abandon farming altogether. But the $3 price floor allows them to continue to grow wheat and remain farmers. So society devotes too many of its scarce resources to wheat production and too few to producing other, more valuable, goods and services. It fails to achieve allocative efficiency.

That's not all. Consumers of wheat-based products pay higher prices because of the price floor. Taxpayers pay higher taxes to finance the government's purchase of the surplus. Also, the price floor causes potential environmental damage by encouraging wheat farmers to bring hilly, erosion-prone "marginal land" into production. The higher price also prompts imports of wheat. But, since such imports would increase the quantity of wheat supplied and thus undermine the price floor, the government needs to erect tariffs (taxes on imports) to keep the foreign wheat out. Such tariffs usually prompt other countries to retaliate with their own tariffs against U.S. agricultural or manufacturing exports.

So it is easy to see why economists "sound the alarm" when politicians advocate imposing price ceilings or price floors such as price controls, rent controls, interest-rate lids, or agricultural price supports. In all these cases, good intentions lead to bad economic outcomes. Government-controlled prices cause shortages or surpluses, distort resource allocation, and produce negative side effects.

FIGURE 3.9 **A price floor.** A price floor is a minimum legal price such as P_f. When the price floor is above the equilibrium price, a persistent product surplus results. Here that surplus is shown by the horizontal distance between Q_s and Q_d.



QUICK REVIEW 3.3

- In competitive markets, prices adjust to the equilibrium level at which quantity demanded equals quantity supplied.
- The equilibrium price and quantity are those indicated by the intersection of the supply and demand curves for any product or resource.
- An increase in demand increases equilibrium price and quantity; a decrease in demand decreases equilibrium price and quantity.
- An increase in supply reduces equilibrium price but increases equilibrium quantity; a decrease in supply increases equilibrium price but reduces equilibrium quantity.
- Over time, equilibrium price and quantity may change in directions that seem at odds with the laws of demand and supply because the other-things-equal assumption is violated.
- Government-controlled prices in the form of ceilings and floors stifle the rationing function of prices, distort resource allocations, and cause negative side effects.

SUMMARY

LO3.1 Characterize and give examples of markets.

Markets bring buyers and sellers together. Some markets are local, others international. Some have physical locations while others are online. For simplicity, this chapter focuses on highly competitive markets in which large numbers of buyers and sellers come together to buy and sell standardized products. All such markets involve demand, supply, price, and quantity, with price being "discovered" through the interacting decisions of buyers and sellers.

LO3.2 Describe *demand* and explain how it can change.

Demand is a schedule or curve representing the willingness of buyers in a specific period to purchase a particular product at each of various prices. The law of demand implies that consumers will buy more of a product at a low price than at a high price. So, other things equal, the relationship between price and quantity demanded is negative or inverse and is graphed as a downsloping curve.

Market demand curves are found by adding horizontally the demand curves of the many individual consumers in the market.

Changes in one or more of the determinants of demand (consumer tastes, the number of buyers in the market, the money incomes of consumers, the prices of related goods, and consumer expectations) shift the market demand curve. A shift to the right is an increase in demand; a shift to the left is a decrease in demand. A change in demand is different from a change in the quantity demanded, the latter being a movement from one point to another point on a fixed demand curve because of a change in the product's price.

LO3.3 Describe *supply* and explain how it can change.

Supply is a schedule or curve showing the amounts of a product that producers are willing to offer in the market at each possible price during a specific period. The law of supply states that, other things equal, producers will offer more of a product at a high price than at a low price. Thus, the relationship between price and quantity supplied is positive or direct, and supply is graphed as an upsloping curve.

The market supply curve is the horizontal summation of the supply curves of the individual producers of the product.

Changes in one or more of the determinants of supply (resource prices, production techniques, taxes or subsidies, the prices of other goods, producer expectations, or the number of sellers in the market) shift the supply curve of a product. A shift to the right is an increase in supply; a shift to the left is a decrease in supply. In contrast, a change in the price of the product being considered causes a change in the quantity supplied, which is shown as a movement from one point to another point on a fixed supply curve.

LO3.4 Relate how supply and demand interact to determine market equilibrium.

The equilibrium price and quantity are established at the intersection of the supply and demand curves. The interaction of market demand and market supply adjusts the price to the point at which the quantities demanded and supplied are equal. This is the equilibrium price. The corresponding quantity is the equilibrium quantity.

The ability of market forces to synchronize selling and buying decisions to eliminate potential surpluses and shortages is known as the rationing function of prices. The equilibrium quantity in competitive markets reflects both productive efficiency (least-cost production) and allocative efficiency (producing the right amount of the product relative to other products).

LO3.5 Explain how changes in supply and demand affect equilibrium prices and quantities.

A change in either demand or supply changes the equilibrium price and quantity. Increases in demand raise both equilibrium price and equilibrium quantity; decreases in demand lower both equilibrium price and equilibrium quantity. Increases in supply lower equilibrium price and raise equilibrium quantity; decreases in supply raise equilibrium price and lower equilibrium quantity.

Simultaneous changes in demand and supply affect equilibrium price and quantity in various ways, depending on their direction and relative magnitudes (see Table 3.3).

LO3.6 Identify what government-set prices are and how they can cause product surpluses and shortages.

A price ceiling is a maximum price set by government and is designed to help consumers. Effective price ceilings produce per sistent product shortages, and if an equitable distribution of the product is sought, government must ration the product to consumers.

A price floor is a minimum price set by government and is designed to aid producers. Effective price floors lead to persistent product surpluses; the government must either purchase the product or eliminate the surplus by imposing restrictions on production or increasing private demand.

Legally fixed prices stifle the rationing function of prices and distort the allocation of resources.

TERMS AND CONCEPTS

demand
demand schedule
law of demand
diminishing marginal utility
income effect
substitution effect
demand curve
determinants of demand
normal goods
inferior goods
substitute good
complementary good

change in demand
change in quantity demanded
supply
supply schedule
law of supply
supply curve
determinants of supply
change in supply
change in quantity supplied
equilibrium price
equilibrium quantity
surplus
shortage
productive efficiency
allocative efficiency
price ceiling
price floor

DISCUSSION QUESTIONS

1. Explain the law of demand. Why does a demand curve slope downward? How is a market demand curve derived from individual demand curves? **LO3.2**
2. What are the determinants of demand? What happens to the demand curve when any of these determinants change? Distinguish between a change in demand and a movement along a fixed demand curve, noting the cause(s) of each. **LO3.2**
3. Explain the law of supply. Why does the supply curve slope upward? How is the market supply curve derived from the supply curves of individual producers? **LO3.3**
4. What are the determinants of supply? What happens to the supply curve when any of these determinants changes? Distinguish between a change in supply and a change in the quantity supplied, noting the cause(s) of each. **LO3.3**
5. In 2001 an outbreak of hoof-and-mouth disease in Europe led to the burning of millions of cattle carcasses. What impact do you think this had on the supply of cattle hides, hide prices, the supply of leather goods, and the price of leather goods? **LO3.5**
6. For each stock in the stock market, the number of shares sold daily equals the number of shares purchased. That is, the quantity of each firm's shares demanded equals the quantity supplied. So, if this equality always occurs, why do the prices of stock shares ever change? **LO3.5**
7. What do economists mean when they say "price floors and ceilings stifle the rationing function of prices and distort resource allocation"? **LO3.6**
8. **LAST WORD** In some countries, such as France, every corpse is available for doctors to "harvest" for organs unless the deceased, while still alive, signed a form forbidding the organs to be harvested. In the United States, it is the opposite: No harvesting is allowed unless the deceased had signed, while still alive, an organ donor form authorizing doctors to harvest any needed organs. Use supply and demand figures to show in which country organ shortages are likely to be less severe.

REVIEW QUESTIONS

1. What effect will each of the following have on the demand for small automobiles such as the Mini-Cooper and Fiat 500? **LO3.2**
   a. Small automobiles become more fashionable.
   b. The price of large automobiles rises (with the price of small autos remaining the same).
   c. Income declines and small autos are an inferior good.
   d. Consumers anticipate that the price of small autos will greatly come down in the near future.
   e. The price of gasoline substantially drops.
2. True or False: A "change in quantity demanded" is a shift of the entire demand curve to the right or to the left. **LO3.2**
3. What effect will each of the following have on the supply of auto tires? **LO3.3**
   a. A technological advance in the methods of producing tires.
   b. A decline in the number of firms in the tire industry.
   c. An increase in the prices of rubber used in the production of tires.
   d. The expectation that the equilibrium price of auto tires will be lower in the future than currently.
   e. A decline in the price of the large tires used for semi trucks and earth-hauling rigs (with no change in the price of auto tires).
   f. The levying of a per-unit tax on each auto tire sold.
   g. The granting of a 50-cent-per-unit subsidy for each auto tire produced.
4. "In the corn market, demand often exceeds supply and supply sometimes exceeds demand." "The price of corn rises and falls in response to changes in supply and demand." In which of these two statements are the terms "supply" and "demand" used correctly? Explain. **LO3.3**
5. Suppose that in the market for computer memory chips, the equilibrium price is $50 per chip. If the current price is $55 per chip, then there will be ______________ of memory chips. **LO3.4**
   a. A shortage.
   b. A surplus.
   c. An equilibrium quantity.
   d None of the above.

6. Critically evaluate: "In comparing the two equilibrium positions in Figure 3.7b, I note that a smaller amount is actually demanded at a lower price. This refutes the law of demand." **LO3.5**
7. Label each of the following scenarios with the set of symbols that best indicates the price change and quantity change that occur in the scenario. In some scenarios, it may not be possible from the information given to determine the direction of a particular price change or a particular quantity change. We will symbolize those cases as, respectively, "P?" and "Q?" The four possible combinations of price and quantity changes are:. **LO3.5**

   P↓ Q?   P? Q↓
   P↑ Q?   P? Q↑

   a. On a hot day, both the demand for lemonade and the supply of lemonade increase.
   b. On a cold day, both the demand for ice cream and the supply of ice cream decrease.
   c. When Hawaii's Mt. Kilauea erupts violently, the demand on the part of tourists for sightseeing flights increases but the supply of pilots willing to provide these dangerous flights decreases.
   d. In a hot area of Arizona where they generate a lot of their electricity with wind turbines, the demand for electricity falls on windy days as people switch off their air conditioners and enjoy the breeze. But at the same time, the amount of electricity supplied increases as the wind turbines spin faster.
8. Suppose the total demand for wheat and the total supply of wheat per month in the Kansas City grain market are as shown in the table below. Suppose that the government establishes a price ceiling of $3.70 for wheat. What might prompt the government to establish this price ceiling? Explain carefully the main effects. Demonstrate your answer graphically. Next, suppose that the government establishes a price floor of $4.60 for wheat. What will be the main effects of this price floor? Demonstrate your answer graphically. **LO3.6**

Thousands of Bushels Demanded	Price per Bushel	Thousands of Bushels Supplied
85	$3.40	72
80	3.70	73
75	4.00	75
70	4.30	77
65	4.60	79
60	4.90	81

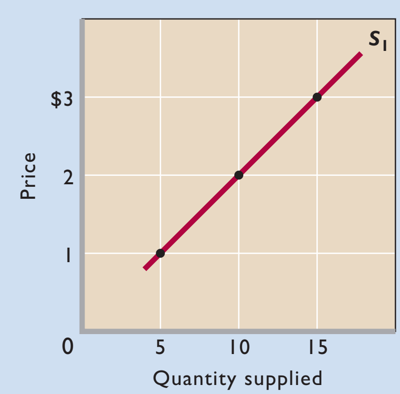
9. A price ceiling will result in a shortage only if the ceiling price is __________ the equilibrium price. **LO3.6**
   a. Less than.
   b. Equal to.
   c. Greater than.
   d. Louder than.

PROBLEMS

1. Suppose there are three buyers of candy in a market: Tex, Dex, and Rex. The market demand and the individual demands of Tex, Dex, and Rex are shown on the next page. **LO3.2**
   a. Fill in the table for the missing values.
   b. Which buyer demands the least at a price of $5? The most at a price of $7?
   c. Which buyer's quantity demanded increases the most when the price is lowered from $7 to $6?
   d. Which direction would the market demand curve shift if Tex withdrew from the market? What if Dex doubled his purchases at each possible price?
   e. Suppose that at a price of $6, the total quantity demanded increases from 19 to 38. Is this a "change in the quantity demanded" or a "change in demand"?

Price per Candy	Individual Quantities Demanded						Total Quantity Demanded
	Tex		Dex		Rex		
$8	3	+	1	+	0	=	___
7	8	+	2	+	___	=	12
6	___	+	3	+	4	=	19
5	17	+	___	+	6	=	27
4	23	+	5	+	8	=	___

2. The figure on the right shows the supply curve for tennis balls, S_1, for Drop Volley Tennis, a producer of tennis equipment. Use the figure and the table below to give your answers to the following questions. **LO3.3**



   a. Use the figure to fill in the quantity supplied on supply curve S_1 for each price in the table below.

Price	S_1 Quantity Supplied	S_2 Quantity Supplied	Change in Quantity Supplied
$3	______	4	______
2	______	2	______
1	______	0	______

b. If production costs were to increase, the quantities supplied at each price would be as shown by the third column of the table ("S_2 Quantity Supplied"). Use those data to draw supply curve S_2 on the same graph as supply curve S_1.

c. In the fourth column of the table, enter the amount by which the quantity supplied at each price changes due to the increase in product costs. (Use positive numbers for increases and negative numbers for decreases.)

d. Did the increase in production costs cause a "decrease in supply" or a "decrease in quantity supplied"?

3. Refer to the expanded table below from review question 8. **LO3.4**

a. What is the equilibrium price? At what price is there neither a shortage nor a surplus? Fill in the surplus-shortage column and use it to confirm your answers.

b. Graph the demand for wheat and the supply of wheat. Be sure to label the axes of your graph correctly. Label equilibrium price P and equilibrium quantity Q.

c. How big is the surplus or shortage at \$3.40? At \$4.90? How big a surplus or shortage results if the price is 60 cents higher than the equilibrium price? 30 cents lower than the equilibrium price?

Thousands of Bushels Demanded	Price per Bushel	Thousands of Bushels Supplied	Surplus (+) or Shortage (−)
85	\$3.40	72	______
80	3.70	73	______
75	4.00	75	______
70	4.30	77	______
65	4.60	79	______
60	4.90	81	______

4. How will each of the following changes in demand and/or supply affect equilibrium price and equilibrium quantity in a competitive market; that is, do price and quantity rise, fall, or remain unchanged, or are the answers indeterminate because they depend on the magnitudes of the shifts? Use supply and demand to verify your answers. **LO3.5**

a. Supply decreases and demand is constant.
b. Demand decreases and supply is constant.
c. Supply increases and demand is constant.
d. Demand increases and supply increases.
e. Demand increases and supply is constant.
f. Supply increases and demand decreases.
g. Demand increases and supply decreases.
h. Demand decreases and supply decreases.

5. Use two market diagrams to explain how an increase in state subsidies to public colleges might affect tuition and enrollments in both public and private colleges. **LO3.5**

6. **ADVANCED ANALYSIS** Assume that demand for a commodity is represented by the equation $P = 10 - .2Q_d$ and supply by the equation $P = 2 + .2Q_s$, where Q_d and Q_s are quantity demanded and quantity supplied, respectively, and P is price. Using the equilibrium condition $Q_s = Q_d$, solve the equations to determine equilibrium price. Now determine equilibrium quantity. **LO3.5**

7. Suppose that the demand and supply schedules for rental apartments in the city of Gotham are as given in the table below. **LO3.6**

Monthly Rent	Apartments Demanded	Apartments Supplied
\$2,500	10,000	15,000
2,000	12,500	12,500
1,500	15,000	10,000
1,000	17,500	7,500
500	20,000	5,000

a. What is the market equilibrium rental price per month and the market equilibrium number of apartments demanded and supplied?

b. If the local government can enforce a rent-control law that sets the maximum monthly rent at \$1,500, will there be a surplus or a shortage? Of how many units? And how many units will actually be rented each month?

c. Suppose that a new government is elected that wants to keep out the poor. It declares that the minimum rent that can be charged is \$2,500 per month. If the government can enforce that price floor, will there be a surplus or a shortage? Of how many units? And how many units will actually be rented each month?

d. Suppose that the government wishes to decrease the market equilibrium monthly rent by increasing the supply of housing. Assuming that demand remains unchanged, by how many units of housing would the government have to increase the supply of housing in order to get the market equilibrium rental price to fall to \$1,500 per month? To \$1,000 per month? To \$500 per month?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

Additional Examples of Supply and Demand

LO3.7 Illustrate how supply and demand analysis can provide insights on actual-economy situations.

Our discussion has clearly demonstrated that supply and demand analysis is a powerful tool for understanding equilibrium prices and quantities. The information provided in the main body of this chapter is fully sufficient for moving forward in the book, but you may find that additional examples of supply and demand are helpful. This optional appendix provides several concrete illustrations of changes in supply and demand.

Your instructor may assign all, some, or none of this appendix, depending on time availability and personal preference.

Changes in Supply and Demand

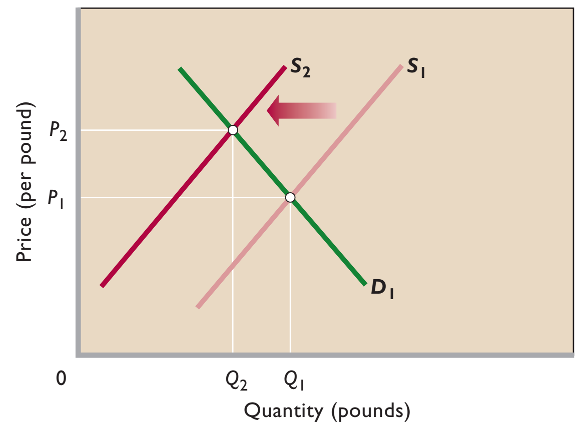
As Figure 3.7 of this chapter demonstrates, changes in supply and demand cause changes in price, quantity, or both. The following applications illustrate this fact in several real-world markets. The simplest situations are those in which either supply changes while demand remains constant or demand changes while supply remains constant. Let's consider two such simple cases first, before looking at more complex applications.

Lettuce

Every now and then we hear on the news that extreme weather has severely reduced the size of some crop. Suppose, for example, that a severe freeze destroys a sizable portion of the lettuce crop. This unfortunate situation implies a significant decline in supply, which we represent as a leftward shift of the supply curve from S_1 to S_2 in Figure 1. At each price, consumers desire as much lettuce as before, so the freeze does not affect the demand for lettuce. That is, demand curve D_1 does not shift.

What are the consequences of the reduced supply of lettuce for equilibrium price and quantity? As shown in Figure 1, the leftward shift of the supply curve disrupts the previous equilibrium in the market for lettuce and drives the equilibrium price upward from P_1 to P_2. Consumers respond to that price hike by reducing the quantity of lettuce demanded from Q_1 to Q_2. Equilibrium is restored at P_2 and Q_2.

FIGURE 1 The market for lettuce. The decrease in the supply of lettuce, shown here by the shift from S_1 to S_2, increases the equilibrium price of lettuce from P_1 to P_2 and reduces the equilibrium quantity from Q_1 to Q_2.



Consumers who are willing and able to pay price P_2 obtain lettuce; consumers unwilling or unable to pay that price do not. Some consumers continue to buy as much lettuce as before, even at the higher price. Others buy some lettuce but not as much as before, and still others opt out of the market completely. The latter two groups use the money they would have spent on lettuce to obtain other products, say, carrots. (Because of our other-things-equal assumption, the prices of other products have not changed.)

Exchange Rates

Exchange rates are the prices at which one currency can be traded (exchanged) for another. Exchange rates are normally determined in foreign exchange markets. One of the largest foreign exchange markets is the euro-dollar market in which the currency used in most of Europe, the *euro*, is exchanged for U.S. dollars. In the United States, this market is set up so that euros are priced in dollars—that is, the "product" being traded is euros and the "price" to buy that product is quoted in dollars. Thus, the market equilibrium price one day might be \$1.25 to buy 1 euro, while on another day it might be \$1.50 to buy 1 euro.

Foreign exchange markets are used by individuals and companies that need to make purchases or payments in a different currency. U.S. companies exporting goods to Germany, for instance, wish to be paid in U.S. dollars. Thus, their German customers will need to convert euros into dollars. The euros that they bring to the euro-dollar market will become part of the overall market supply of euros. Conversely, an American mutual fund may wish to purchase some French real estate outside of Paris. But to purchase that real estate, it will need to pay in euros because the current French owners will only accept payment in euros. Thus, the American mutual fund has a demand to purchase euros that will form part of the overall market demand for euros. The fund will bring dollars to the euro-dollar foreign exchange market in order to purchase the euros it desires.

Sometimes, the demand for euros increases. This might be because a European product surges in popularity in foreign countries. For example, if a new German-made automobile is a big hit in the United States, American car dealers will demand more euros with which to pay for more units of that new model. This will shift the demand curve for euros to the right, as from D_1 to D_2 in Figure 2. Given the fixed euro supply curve S_1, the increase in demand raises the equilibrium exchange rate (the equilibrium number of dollars needed to purchase 1 euro) from \$1.25 to \$1.50. The equilibrium quantity of euros purchased increases from Q_1 to Q_2. Because a higher dollar amount is now needed to purchase one euro, economists say that the dollar has *depreciated*—gone down in value—relative to the euro. Alternatively, the euro has *appreciated*—gone up in value—relative to the dollar, because one euro now buys \$1.50 rather than \$1.25.

Pink Salmon

Now let's see what happens when both supply and demand change at the same time. Several decades ago, people who caught salmon earned as much as \$1 for each pound of pink salmon—the type of salmon most commonly used for canning. In Figure 3 that price is represented as P_1, at the intersection of supply curve S_1 and demand curve D_1. The corresponding quantity of pink salmon is shown as Q_1 pounds.

As time passed, supply and demand changed in the market for pink salmon. On the supply side, improved technology in the form of larger, more efficient fishing boats greatly increased the catch and lowered the cost of obtaining it. Also, high profits at price P_1 encouraged many new fishers to enter the industry. As a result of these changes, the supply of pink salmon greatly increased and the supply curve shifted to the right, as from S_1 to S_2 in Figure 3.

Over the same years, the demand for pink salmon declined, as represented by the leftward shift from D_1 to D_2 in Figure 3. That decrease was caused by increases in consumer income and reductions of the price of substitute products. As buyers' incomes rose, consumers shifted demand away from canned fish and toward higher-quality fresh or frozen fish, including more-valued Atlantic, chinook, sockeye, and coho salmon. Moreover, the emergence of fish farming, in which salmon are raised in ocean

FIGURE 2 The market for euros. The increase in the demand for euros, shown here by the shift from D_1 to D_2, increases the equilibrium price of one euro from \$1.25 to \$1.50 and increases the equilibrium quantity of euros that are exchanged from Q_1 to Q_2. The dollar has depreciated.

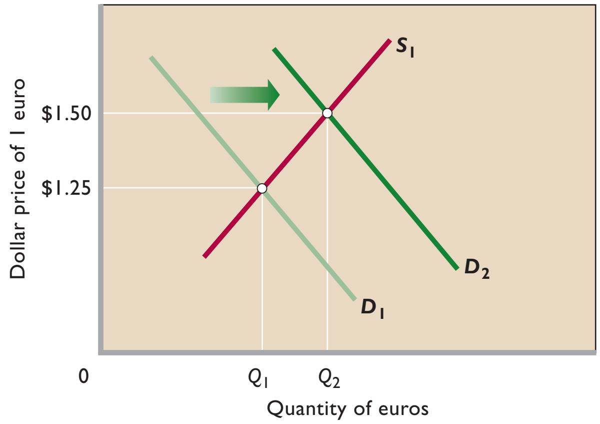
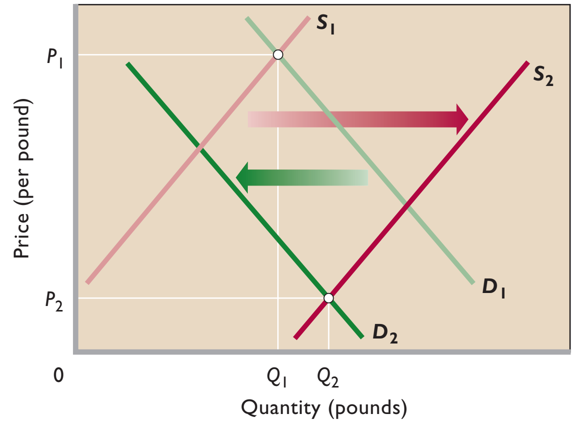


FIGURE 3 The market for pink salmon. In the last several decades, the supply of pink salmon has increased and the demand for pink salmon has decreased. As a result, the price of pink salmon has declined, as from P_1 to P_2. Because supply has increased by more than demand has decreased, the equilibrium quantity of pink salmon has increased, as from Q_1 to Q_2.



net pens, lowered the prices of these substitute species. That, too, reduced the demand for pink salmon.

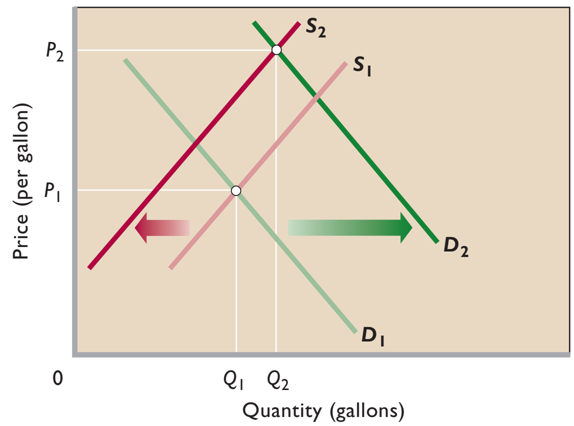
The altered supply and demand reduced the price of pink salmon to as low as \$0.10 per pound, as represented by the drop in price from P_1 to P_2 in Figure 3. Both the supply increase and the demand decrease helped reduce the equilibrium price. However, in this particular case the equilibrium quantity of pink salmon increased, as represented by the move from Q_1 to Q_2. Both shifts reduced the equilibrium price, but equilibrium quantity increased because the increase in supply exceeded the decrease in demand.

Gasoline

The price of gasoline in the United States has increased rapidly several times during the past several years. For example, the average price of a gallon of gasoline rose from around \$2.60 in October 2010 to about \$3.90 in May 2011. What caused this 50 percent rise in the price of gasoline? How would we diagram this increase?

We begin in Figure 4 with the price of a gallon of gasoline at P_1, representing the \$2.60 price. Simultaneous supply and demand factors disturbed this equilibrium. Supply uncertainties relating to Middle East politics and warfare and expanded demand for oil by fast-growing countries such as China pushed up the price of a barrel of oil from under \$80 per barrel in October 2010 to well over \$100 per barrel in May 2011. Oil is the main input for producing gasoline, so any sustained rise in its price boosts the per-unit cost of producing gasoline. Such cost rises decrease the supply of gasoline, as represented by the leftward shift of the supply curve from S_1 to S_2 in Figure 4. At times refinery breakdowns in the United States also contributed to this reduced supply.

FIGURE 4 **The market for gasoline.** An increase in the demand for gasoline, as shown by the shift from D_1 to D_2, coupled with a decrease in supply, as shown by the shift from S_1 to S_2, boosts equilibrium price (here from P_1 to P_2). In this case, equilibrium quantity increases from Q_1 to Q_2 because the increase in demand outweighs the decrease in supply.



While the supply of gasoline declined between October 2010 and May 2011, the demand for gasoline increased, as depicted by the rightward shift of the demand curve from D_1 to D_2. Incomes in general were rising over this period because the U.S. economy was expanding. Rising incomes raise demand for all normal goods, including gasoline. An increased number of low-gas-mileage SUVs and light trucks on the road also contributed to growing gas demand.

The combined decline in gasoline supply and increase in gasoline demand boosted the price of gasoline from \$2.60 to \$3.90, as represented by the rise from P_1 to P_2 in Figure 4. Because the demand increase outweighed the supply decrease, the equilibrium quantity expanded, here from Q_1 to Q_2.

In other periods the price of gasoline has *declined* as the demand for gasoline has increased. Test your understanding of the analysis by explaining how such a price decrease could occur.

Sushi

Sushi bars are springing up like Starbucks in American cities (well, maybe not that fast!). Consumption of sushi, the raw-fish delicacy from Japan, has soared in the United States in recent years. Nevertheless, the price of sushi has remained relatively constant.

Supply and demand analysis helps explain this circumstance of increased quantity and constant price. A change in tastes has increased the U.S. demand for sushi. Many consumers of sushi find it highly tasty when they try it. And, as implied by the growing number of sushi bars in the United States, the supply of sushi has also expanded.

We represent these supply and demand changes in Figure 5 as the rightward shift of the demand curve from D_1 to D_2 and the rightward shift of the supply curve from S_1 to S_2. Observe that the equilibrium quantity of sushi increases from Q_1 to Q_2 and equilibrium price remains constant at P_1. The increase in supply, which taken alone would reduce price, has perfectly offset the increase in demand, which taken alone would raise price. The price of sushi does not change, but the equilibrium quantity greatly increases because both the increase in demand and the increase in supply expand purchases and sales.

Simultaneous increases in demand and supply can cause price to either rise, fall, or remain constant, depending on the relative magnitudes of the supply and demand increases. In this case, price remained constant.

FIGURE 5 The market for sushi. Equal increases in the demand for sushi, as from D_1 to D_2, and in the supply of sushi, as from S_1 to S_2, expand the equilibrium quantity of sushi (here from Q_1 to Q_2) while leaving the price of sushi unchanged at P_1.

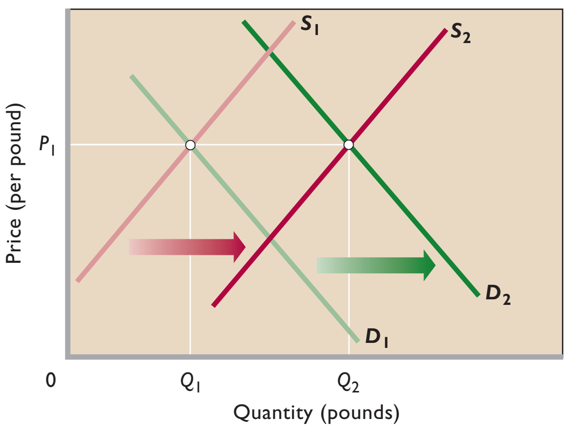
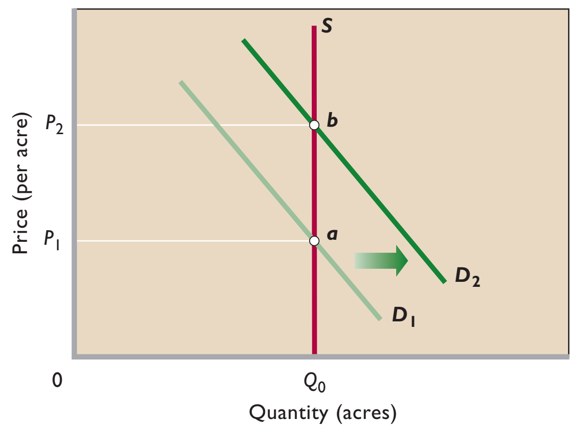


FIGURE 6 The market for land in San Francisco. Because the quantity of land in San Francisco is fixed at Q_0, the supply curve is vertical above Q_0 in order to indicate that the same quantity of land will be supplied no matter what the price is. As demand increases from D_1 to D_2, the equilibrium price rises from P_1 to P_2. Because the quantity of land is fixed at Q_0, the movement from equilibrium *a* to equilibrium *b* involves only a change in the equilibrium price; the equilibrium quantity remains at Q_0 due to land being in fixed supply.



Upsloping versus Vertical Supply Curves

As you already know, the typical good or service possesses an upsloping supply curve because a higher market price will cause producers to increase the quantity supplied. There are, however, some goods and services whose quantities supplied are fixed and totally unresponsive to changes in price. Examples include the amount of land in a given area, the number of seats in a stadium, and the limited part of the electromagnetic spectrum that is reserved for cellular telephone transmissions. These sorts of goods and services have vertical supply curves because the same fixed amount is available no matter what price is offered to suppliers.

Reactions to Demand Shifts

Markets react very differently to a shift in demand depending upon whether they have upsloping or vertical supply curves.

Upsloping Supply Curves When a market has an upsloping supply curve, any shift in demand will cause both the equilibrium price *and* the equilibrium quantity to adjust. Consider Figure 2. When the demand for euros increases, the movement from the initial equilibrium to the final equilibrium involves the equilibrium price rising from $1.25 to $1.50 while the equilibrium quantity increases from Q_1 to Q_2. Price and quantity *both* change.

Vertical Supply Curves When a market has a vertical supply curve, any shift in demand will cause only the equilibrium price to change; the equilibrium quantity remains the same because the quantity supplied is fixed and cannot adjust.

Consider Figure 6, in which the supply of land in San Francisco is fixed at quantity Q_0. If demand increases from D_1 to D_2, the movement from the initial equilibrium at point *a* to the final equilibrium at point *b* is accomplished solely by a rise in the equilibrium price from P_1 to P_2. Because the quantity of land is fixed, the increase in demand cannot cause any change in the equilibrium quantity supplied. The entire adjustment from the initial equilibrium to the final equilibrium has to come in the form of a higher equilibrium price.

This fact explains why real estate prices are so high in San Francisco and other major cities. Any increase in demand cannot be met by a combination of increases in price and increases in quantity. With the quantity of land in fixed supply, any increase in demand results solely in higher equilibrium land prices.

Preset Prices

In the body of this chapter, we saw that an effective government-imposed price ceiling (legal maximum price) causes quantity demanded to exceed quantity supplied—a shortage. An effective government-imposed price floor (legal minimum price) causes quantity supplied to exceed quantity demanded—a surplus. Put simply: Shortages result when prices are set below, and surpluses result when prices are set above, equilibrium prices.

We now want to establish that shortages and surpluses can occur in markets other than those in which government imposes price floors and ceilings. Such market imbalances happen when the seller or sellers set prices in advance of sales and the prices selected turn out to be below or above equilibrium prices. Consider the following two examples.

Olympic Figure Skating Finals

Tickets for the women's figure skating championship at the Olympics are among the world's "hottest tickets." The popularity of this event and the high incomes of buyers translate into tremendous ticket demand. The Olympic officials set the price for the tickets in advance. Invariably, the price, although high, is considerably below the equilibrium price that would equate quantity demanded and quantity supplied. A severe shortage of tickets therefore occurs in this *primary market*—the market involving the official ticket office.

The shortage, in turn, creates a *secondary market* in which buyers bid for tickets held by initial purchasers rather than the original seller. Scalping tickets—selling them above the original ticket price—may be legal or illegal, depending on local laws.

Figure 7 shows how the shortage in the primary ticket market looks in terms of supply and demand analysis. Demand curve *D* represents the strong demand for tickets and supply curve *S* represents the supply of tickets. The supply curve is vertical because a fixed number of tickets are printed to match the capacity of the arena. At the printed ticket price of P_1, the quantity of tickets demanded, Q_2, exceeds the quantity supplied, Q_1. The result is a shortage of *ab*—the horizontal distance between Q_2 and Q_1 in the primary market.

If the printed ticket price had been the higher equilibrium price P_2, no shortage of tickets would have occurred. But at the lower price P_1, a shortage and secondary ticket market will emerge among those buyers willing to pay more than the printed ticket price and those sellers willing to sell their purchased tickets for more than the original price. Wherever there are shortages and secondary markets, it is safe to assume the original price was set below the equilibrium price.

Olympic Curling Preliminaries

Contrast the shortage of tickets for the women's figure skating finals at the Olympics to the surplus of tickets for one of the preliminary curling matches. For the uninitiated, curling is a sport in which participants slide a heavy round object called a "stone" down the ice toward a target while teammates called "sweepers" use brooms to alter the course of the stone when desired.

Curling is a popular spectator sport in a few nations such as Canada, but it does not draw many fans in most countries. So the demand for tickets to most of the preliminary curling events is not very strong. We demonstrate this weak demand as *D* in Figure 8. As in our previous

FIGURE 7 The market for tickets to the Olympic women's figure skating finals. The demand curve *D* and supply curve *S* for the Olympic women's figure skating finals produce an equilibrium price that is above the P_1 price printed on the ticket. At price P_1 the quantity of tickets demanded, Q_2, greatly exceeds the quantity of tickets available, Q_1. The resulting shortage of *ab* (= $Q_2 - Q_1$) gives rise to a legal or illegal secondary market.

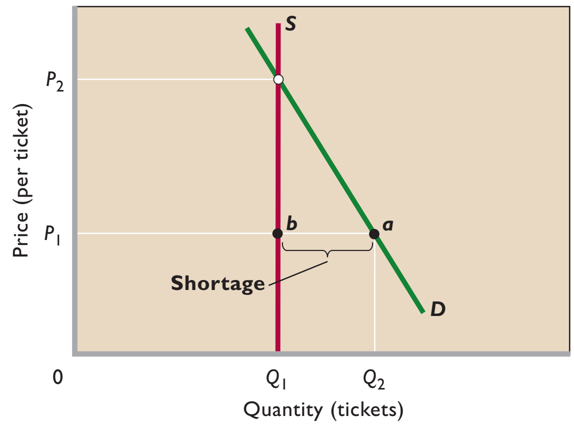
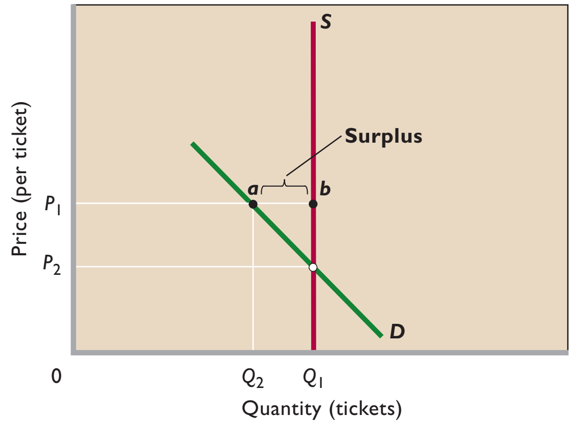


FIGURE 8 The market for tickets to the Olympic curling preliminaries. The demand curve *D* and supply curve *S* for the Olympic curling preliminaries produce an equilibrium price below the P_1 price printed on the ticket. At price P_1 the quantity of tickets demanded is less than the quantity of tickets available. The resulting surplus of *ba* (= $Q_1 - Q_2$) means the event is not sold out.



example, the supply of tickets is fixed by the size of the arena and is shown as vertical line *S*.

We represent the printed ticket price as P_1 in Figure 8. In this case the printed price is much higher than the equilibrium price of P_2. At the printed ticket price, quantity supplied is Q_1 and quantity demanded is Q_2. So a surplus of tickets of *ba* ($= Q_1 - Q_2$) occurs. No ticket scalping occurs and there are numerous empty seats. Only if the Olympic officials had priced the tickets at the lower price P_2 would the event have been a sellout. (Actually, the Olympic officials try to adjust to demand realities for curling contests by holding them in smaller arenas and by charging less for tickets. Nevertheless, the stands are rarely full for the preliminary contests, which compete against final events in other winter Olympic sports.)

APPENDIX SUMMARY

LO3.7 Illustrate how supply and demand analysis can provide insights on actual-economy situations.

A decrease in the supply of a product increases its equilibrium price and reduces its equilibrium quantity. In contrast, an increase in the demand for a product boosts both its equilibrium price and its equilibrium quantity.

Simultaneous changes in supply and demand affect equilibrium price and quantity in various ways, depending on the relative magnitudes of the changes in supply and demand. Equal increases in supply and demand, for example, leave equilibrium price unchanged.

Products (such as land) whose quantities supplied do not vary with price have vertical supply curves. For these products, any shift in demand will lead to a change in the equilibrium price but no change in the equilibrium quantity.

Sellers set prices of some items such as tickets in advance of the event. These items are sold in the primary market that involves the original sellers and buyers. If preset prices turn out to be below the equilibrium prices, shortages occur and scalping in legal or illegal secondary markets arises. The prices in the secondary market then rise above the preset prices. In contrast, surpluses occur when the preset prices happen to exceed the equilibrium prices.

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Why are shortages or surpluses more likely with preset prices, such as those on tickets, than flexible prices, such as those on gasoline? **LO3.7**
2. Most scalping laws make it illegal to sell—but not to buy—tickets at prices above those printed on the tickets. Assuming that is the case, use supply and demand analysis to explain why the equilibrium ticket price in an illegal secondary market tends to be higher than in a legal secondary market. **LO3.7**
3. Go to the Web site of the Energy Information Administration, **www.eia.doe.gov**, and follow the links to find the current retail price of gasoline. How does the current price of regular gasoline compare with the price a year ago? What must have happened to either supply, demand, or both to explain the observed price change? **LO3.7**
4. Suppose the supply of apples sharply increases because of perfect weather conditions throughout the growing season. Assuming no change in demand, explain the effect on the equilibrium price and quantity of apples. Explain why quantity demanded increases even though demand does not change. **LO3.7**
5. Assume the demand for lumber suddenly rises because of a rapid growth of demand for new housing. Assume no change in supply. Why does the equilibrium price of lumber rise? What would happen if the price did not rise under the demand and supply circumstances described? **LO3.7**
6. Assume that both the supply of bottled water and the demand for bottled water rise during the summer but that supply increases more rapidly than demand. What can you conclude about the directions of the impacts on equilibrium price and equilibrium quantity? **LO3.7**
7. When asked for investment advice, humorist Will Rogers joked that people should "[b]uy land. They ain't making any more of the stuff." Explain his advice in terms of the supply and demand model. **LO3.7**

APPENDIX REVIEW QUESTIONS

1. Will the equilibrium price of orange juice increase or decrease in each of the following situations? **LO3.7**
   a. A medical study reporting that orange juice reduces cancer is released at the same time that a freak storm destroys half of the orange crop in Florida.
   b. The prices of all beverages except orange juice fall by half while unexpectedly perfect weather in Florida results in an orange crop that is 20 percent larger than normal.
2. Consider the market for coffee beans. Suppose that the prices of all other caffeinated beverages go up 30 percent while at the same time a new fertilizer boosts production at coffee plantations dramatically. Which of the following best describes what is likely to happen to the equilibrium price and quantity of coffee beans? **LO3.7**
   a. Both the equilibrium price and the quantity will rise.
   b. The equilibrium price will rise but the equilibrium quantity will fall.
   c. The equilibrium price may rise or fall but the equilibrium quantity will rise for certain.
   d. Neither the price change nor the quantity change can be determined for certain.
   e. None of the above.
3. A price ceiling will result in a shortage only if the ceiling price is __________ the equilibrium price. **LO3.7**
   a. Less than.
   b. Equal to.
   c. Greater than.
   d. Faster than.
4. Suppose that you are the economic advisor to a local government that has to deal with a politically embarrassing surplus that was caused by a price floor that the government recently imposed. Your first suggestion is to get rid of the price floor, but the politicians don't want to do that. Instead, they present you with the following list of options that they hope will get rid of the surplus while keeping the price floor. Identify each one as either *could work* or *can't work*. **LO3.7**
   a. Restricting supply.
   b. Decreasing demand.
   c. Purchasing the surplus at the floor price.
5. Suppose both the demand for olives and the supply of olives decline by equal amounts over some time period. Use graphical analysis to show the effect on equilibrium price and quantity. **LO3.7**
6. Governments can use subsidies to increase demand. For instance, a government can pay farmers to use organic fertilizers rather than traditional fertilizers. That subsidy increases the demand for organic fertilizer. Consider two industries, one in which supply is nearly vertical and the other in which supply is nearly horizontal. Assume that firms in both industries would prefer a higher market equilibrium price because a higher market equilibrium price would mean higher profits. Which industry would probably spend more resources lobbying the government to increase the demand for its output? (Assume that both industries have similarly sloped demand curves.) **LO3.7**
   a. The industry with a nearly flat supply curve.
   b. The industry with a nearly vertical supply curve.

APPENDIX PROBLEMS

1. Demand and supply often shift in the retail market for gasoline. Here are two demand curves and two supply curves for gallons of gasoline in the month of May in a small town in Maine. Some of the data are missing. **LO3.7**

	Quantities Demanded		Quantities Supplied	
Price	D_1	D_2	S_1	S_2
$4.00	5,000	7,500	9,000	9,500
____	6,000	8,000	8,000	9,000
2.00	____	8,500	____	8,500
____	____	9,000	5,000	____

   a. Use the following facts to fill in the missing data in the table. If demand is D_1 and supply is S_1, the equilibrium quantity is 7,000 gallons per month. When demand is D_2 and supply is S_1, the equilibrium price is $3.00 per gallon. When demand is D_2 and supply is S_1, there is an excess demand of 4,000 gallons per month at a price of $1.00 per gallon. If demand is D_1 and supply is S_2, the equilibrium quantity is 8,000 gallons per month.
   b. Compare two equilibriums. In the first, demand is D_1 and supply is S_1. In the second, demand is D_1 and supply is S_2. By how much does the equilibrium quantity change? By how much does the equilibrium price change?
   c. If supply falls from S_2 to S_1 while demand declines from D_2 to D_1, does the equilibrium price rise, fall, or stay the same? What if only supply falls? What if only demand falls?
   d. Suppose that supply is fixed at S_1 and that demand starts at D_1. By how many gallons per month would demand have to increase at each price level such that the equilibrium price per gallon would be $3.00? $4.00?
2. The table at the top of the next page shows two demand schedules for a given style of men's shoe—that is, how many pairs per month will be demanded at various prices at a men's clothing store in Seattle called Stromnord.

Price	D_1 Quantity Demanded	D_2 Quantity Demanded
$75	53	13
70	60	15
65	68	18
60	77	22
55	87	27

Suppose that Stromnord has exactly 65 pairs of this style of shoe in inventory at the start of the month of July and will not receive any more pairs of this style until at least August 1. **LO3.7**

a. If demand is D_1, what is the lowest price that Stromnord can charge so that it will not run out of this model of shoe in the month of July? What if demand is D_2?

b. If the price of shoes is set at $75 for both July and August and demand will be D_2 in July and D_1 in August, how many pairs of shoes should Stromnord order if it wants to end the month of August with exactly zero pairs of shoes in its inventory? What if the price is set at $55 for both months?

3. Use the table below to answer the questions that follow: **LO3.7**

a. If this table reflects the supply of and demand for tickets to a particular World Cup soccer game, what is the stadium capacity?

b. If the preset ticket price is $45, would we expect to see a secondary market for tickets? Would the price of a ticket in the secondary market be higher than, the same as, or lower than the price in the primary (original) market?

c. Suppose for some other World Cup game the quantity of tickets demanded is 20,000 lower at each ticket price than shown in the table. If the ticket price remains $45, would the event be a sellout?

Quantity Demanded, Thousands	Price	Quantity Supplied, Thousands
80	$25	60
75	35	60
70	45	60
65	55	60
60	65	60
55	75	60
50	85	60

CHAPTER **4**

Market Failures: Public Goods and Externalities

Learning Objectives

LO4.1 Differentiate between demand-side market failures and supply-side market failures.

LO4.2 Explain the origin of both consumer surplus and producer surplus, and explain how properly functioning markets maximize their sum, total surplus, while optimally allocating resources.

LO4.3 Describe free riding and public goods, and illustrate why private firms cannot normally produce public goods.

LO4.4 Explain how positive and negative externalities cause under- and overallocations of resources.

LO4.5 Show why we normally won't want to pay what it would cost to eliminate every last bit of a negative externality such as air pollution.

LO4.6 (Appendix) Describe how information failures may justify government intervention in some markets.

Competitive markets usually do a remarkably effective job of allocating society's scarce resources to their most highly valued uses. Thus, we begin this chapter by demonstrating how properly functioning markets efficiently allocate resources. We then explore what happens when markets don't function properly. In some circumstances, economically desirable goods are not produced at all. In other situations, they are either overproduced or underproduced. This chapter focuses on these situations, which economists refer to as **market failures.**

In such situations, an economic role for government may arise. We will examine that role as it relates to public goods and so-called externalities—situations where market failures lead to suboptimal outcomes that the government may be able to improve upon by using its powers to tax, spend, and regulate. The government may, for instance, pay for the production of goods that the private sector fails to produce. It may also act to reduce the production of those goods and services that the private sector overproduces. Implementing such policies can, however, be both costly and complicated. Thus, we conclude the chapter by noting the government inefficiencies that can hinder government efforts to improve economic outcomes.

Market Failures in Competitive Markets[1]

LO4.1 Differentiate between demand-side market failures and supply-side market failures.

In Chapter 3 we asserted that "competitive markets usually produce an assignment of resources that is 'right' from an economic perspective." We now want to focus on the word "usually" and discuss exceptions. We must do this because it is unfortunately the case that the presence of robust competition involving many buyers and many sellers may not, by itself, be enough to guarantee that a market will allocate resources correctly. Market failures sometimes happen in competitive markets. The focus of this chapter is to explain how and why such market failures can arise.

Fortunately, the broad picture is simple. Market failures in competitive markets fall into just two categories:

- **Demand-side market failures** happen when demand curves do not reflect consumers' full willingness to pay for a good or service.
- **Supply-side market failures** occur when supply curves do not reflect the full cost of producing a good or service.

Demand-Side Market Failures

Demand-side market failures arise because it is impossible in certain cases to charge consumers what they are willing to pay for a product. Consider outdoor fireworks displays. People enjoy fireworks and would therefore be *willing* to pay to see a fireworks display if the only way to see it was to have to pay for the right to do so. But because such displays are outdoors and in public, people don't actually *have* to pay to see the display because there is no way to exclude those who haven't paid from also enjoying the show. Private firms will therefore be unwilling to produce outdoor fireworks displays, as it will be nearly impossible for them to raise enough revenue to cover production costs.

Supply-Side Market Failures

Supply-side market failures arise in situations in which a firm does not have to pay the full cost of producing its output. Consider a coal-burning power plant. The firm running the plant will have to pay for all of the land, labor, capital, and entrepreneurship that it uses to generate electricity by burning coal. But if the firm is not charged for the smoke that it releases into the atmosphere, it will fail to pay another set of costs—the costs that its pollution imposes on other people. These include future harm from global warming, toxins that affect wildlife, and possible damage to agricultural crops downwind.

A market failure arises because it is not possible for the market to correctly weigh costs and benefits in a situation in which some of the costs are completely unaccounted for. The coal-burning power plant produces more electricity and generates more pollution than it would if it had to pay for each ton of smoke that it released into the atmosphere. The extra units that are produced are units of output for which the costs are *greater than* the benefits. Obviously, these units should not be produced.

[1]Other market failures arise when there are not enough buyers or sellers to ensure competition. In those situations, the lack of competition allows either buyers or sellers to restrict purchases or sales below optimal levels for their own benefit. As an example, a monopoly—a firm that is the only producer in its industry—can restrict the amount of output that it supplies in order to drive up the market price and thereby increase its own profit.

Efficiently Functioning Markets

LO4.2 Explain the origin of both consumer surplus and producer surplus, and explain how properly functioning markets maximize their sum, total surplus, while optimally allocating resources.

The best way to understand market failure is to first understand how properly functioning competitive markets achieve economic efficiency. We touched on this subject in Chapter 3, but we now want to expand and deepen that analysis, both for its own sake and to set up our discussion of public goods and externalities. Two conditions must hold if a competitive market is to produce efficient outcomes: The demand curve in the market must reflect consumers' full willingness to pay, and the supply curve in the market must reflect all the costs of production. If these conditions hold, then the market will produce only units for which benefits are at least equal to costs. It will also maximize the amount of "benefit surpluses" that are shared between consumers and producers.

Consumer Surplus

The benefit surplus received by a consumer or consumers in a market is called **consumer surplus.** It is defined as the difference between the maximum price a consumer is (or consumers are) willing to pay for a product and the actual price that they do pay.

The maximum price that a person is willing to pay for a unit of a product depends on the opportunity cost of that person's consumption alternatives. Suppose that Ted is offered the chance to purchase an apple. He would of course like to have it for free, but the maximum amount he would be willing to pay depends on the alternative uses to which he can put his money. If his maximum willingness to pay for that particular apple is $1.25, then we know that he is willing to forgo up to—but not more than—$1.25 of other goods and services. Paying even one cent more would entail having to give up too much of other goods and services.

It also means that if Ted is charged any market price less than $1.25, he will receive a consumer surplus equal to the difference between the $1.25 maximum price that he would have been willing to pay and the lower market price. For instance, if the market price is $0.50 per apple, Ted will receive a consumer surplus of $0.75 per apple (= $1.25 − $0.50). In nearly all markets, consumers individually and collectively gain greater total utility or satisfaction in dollar terms from their purchases than the amount of their expenditures (= product price × quantity). This utility surplus arises because each consumer who buys the product only has to pay the market equilibrium price

TABLE 4.1 Consumer Surplus

(1) Person	(2) Maximum Price Willing to Pay	(3) Actual Price (Equilibrium Price)	(4) Consumer Surplus
Bob	$13	$8	$5 (= $13 − $8)
Barb	12	8	4 (= $12 − $8)
Bill	11	8	3 (= $11 − $8)
Bart	10	8	2 (= $10 − $8)
Brent	9	8	1 (= $9 − $8)
Betty	8	8	0 (= $8 − $8)

even though many of them would have been willing to pay more than the equilibrium price to obtain the product.

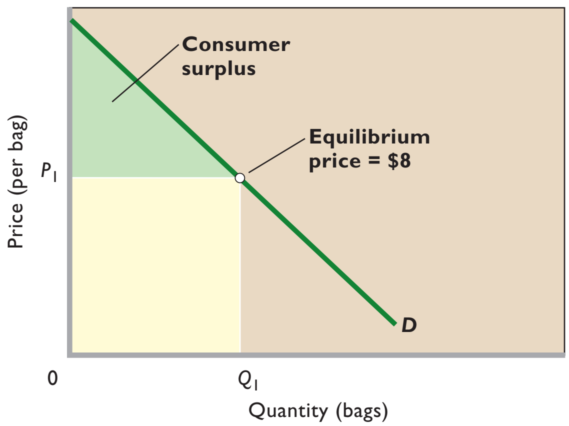
The concept of maximum willingness to pay also gives us another way to understand demand curves. Consider Table 4.1, where the first two columns show the maximum amounts that six consumers would each be willing to pay for a bag of oranges. Bob, for instance, would be willing to pay a maximum of $13 for a bag of oranges. Betty, by contrast, would only be willing to pay a maximum of $8 for a bag of oranges.

Notice that the maximum prices that these individuals are willing to pay represent points on a demand curve because the lower the market price, the more bags of oranges will be demanded. At a price of $12.50, for instance, Bob will be the only person listed in the table who will purchase a bag. But at a price of $11.50, both Bob and Barb will want to purchase a bag. And at a price of $10.50, Bob, Barb, and Bill will each want to purchase a bag. The lower the price, the greater the total quantity demanded as the market price falls below the maximum prices of more and more consumers.

Lower prices also imply larger consumer surpluses. When the price is $12.50, Bob only gets $0.50 in consumer surplus because his maximum willingness to pay of $13 is only $0.50 higher than the market price of $12.50. But if the market price were to fall to $8, then his consumer surplus would be $5 (= $13 − $8). The third and fourth columns of Table 4.1 show how much consumer surplus each of our six consumers will receive if the market price of a bag of oranges is $8. Only Betty receives no consumer surplus because her maximum willingness to pay exactly matches the $8 equilibrium price.

It is easy to show on a graph both the individual consumer surplus received by each particular buyer in a market as well as the collective consumer surplus received by all buyers. Consider Figure 4.1, which shows the market equilibrium price P_1 = $8 as well as the downsloping

FIGURE 4.1 Consumer surplus. Consumer surplus—shown as the green triangle—is the difference between the maximum prices consumers are willing to pay for a product and the lower equilibrium price, here assumed to be \$8. For quantity Q_1, consumers are willing to pay the sum of the amounts represented by the green triangle and the yellow rectangle. Because they need to pay only the amount shown as the yellow rectangle, the green triangle shows consumer surplus.



demand curve *D* for bags of oranges. Demand curve *D* includes not only the six consumers named in Table 4.1 but also every other consumer of oranges in the market. The individual consumer surplus of each particular person who is willing to buy at the \$8 market price is simply the vertical distance from the horizontal line that marks the \$8 market price up to that particular buyer's maximum willingness to pay. The collective consumer surplus obtained by all of our named and unnamed buyers is found by adding together each of their individual consumer surpluses. To obtain the Q_1 bags of oranges represented, consumers collectively are willing to pay the total amount shown by the sum of the green triangle and yellow rectangle under the demand curve and to the left of Q_1. But consumers need pay only the amount represented by the yellow rectangle ($= P_1 \times Q_1$). So the green triangle is the consumer surplus in this market. It is the sum of the vertical distances between the demand curve and the \$8 equilibrium price at each quantity up to Q_1. Alternatively, it is the sum of the gaps between maximum willingness to pay and actual price, such as those we calculated in Table 4.1. Thus, consumer surplus can also be defined as the area that lies below the demand curve and above the price line that extends horizontally from P_1.

Consumer surplus and price are inversely (negatively) related. Given the demand curve, higher prices reduce consumer surplus; lower prices increase it. To test this generalization, draw in an equilibrium price above \$8 in Figure 4.1 and observe the reduced size of the triangle representing consumer surplus. When price goes up, the gap narrows between the maximum willingness to pay and the actual price. Next, draw in an equilibrium price below \$8 and see that consumer surplus increases. When price declines, the gap widens between maximum willingness to pay and actual price.

ORIGIN OF THE IDEA

O4.1

Consumer surplus

Producer Surplus

Like consumers, producers also receive a benefit surplus in markets. This **producer surplus** is the difference between the actual price a producer receives (or producers receive) and the minimum acceptable price that a consumer would have to pay the producer to make a particular unit of output available.

A producer's minimum acceptable price for a particular unit will equal the producer's marginal cost of producing that particular unit. That marginal cost will be the sum of the rent, wages, interest, and profit that the producer will need to pay in order to obtain the land, labor, capital, and entrepreneurship required to produce that particular unit. In this section, we are assuming that the marginal cost of producing a unit will include *all* of the costs of production. Unlike the coal-burning power plant mentioned previously, the producer must pay for all of its costs, including the cost of pollution. In later sections, we will explore the market failures that arise in situations where firms do not have to pay all their costs.

In addition to equaling marginal cost, a producer's minimum acceptable price can also be interpreted as the opportunity cost of bidding resources away from the production of other products. To see why this is true, suppose that Leah is an apple grower. The resources necessary for her to produce one apple could be used to produce other things. To get them directed toward producing an apple, it is necessary to pay Leah what it will cost her to bid the necessary resources away from other entrepreneurs who would like to use them to produce other products. Leah would, naturally, like to get paid as much as possible to produce the apple for you. But her minimum acceptable price is the lowest price you could pay her such that she can just break even after bidding away from other uses the land, labor, capital, and entrepreneurship necessary to produce the apple.

The size of the producer surplus earned on any particular unit will be the difference between the market price that the producer actually receives and the producer's

TABLE 4.2 **Producer Surplus**

(1) Person	(2) Minimum Acceptable Price	(3) Actual Price (Equilibrium Price)	(4) Producer Surplus
Carlos	\$3	\$8	\$5 (= \$8 − \$3)
Courtney	4	8	4 (= \$8 − \$4)
Chuck	5	8	3 (= \$8 − \$5)
Cindy	6	8	2 (= \$8 − \$6)
Craig	7	8	1 (= \$8 − \$7)
Chad	8	8	0 (= \$8 − \$8)

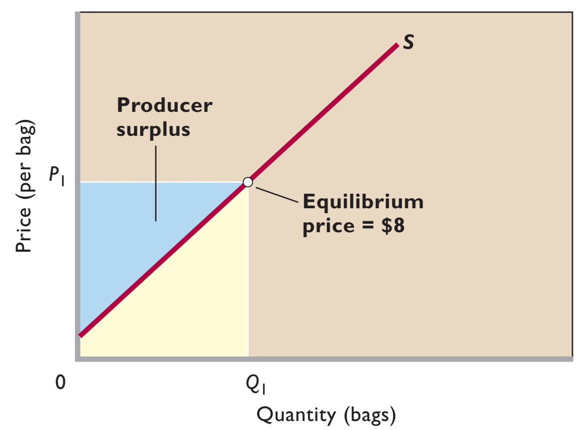
minimum acceptable price. Consider Table 4.2, which shows the minimum acceptable prices of six different orange growers. With a market price of \$8, Carlos, for instance, has a producer surplus of \$5, which is equal to the market price of \$8 minus his minimum acceptable price of \$3. Chad, by contrast, receives no producer surplus because his minimum acceptable price of \$8 just equals the market equilibrium price of \$8.

Carlos's minimum acceptable price is lower than Chad's minimum acceptable price because Carlos is a more efficient producer than Chad, by which we mean that Carlos produces oranges using a less-costly combination of resources than Chad uses. The differences in efficiency between Carlos and Chad are likely due to differences in the type and quality of resources available to them. Carlos, for instance, may own land perfectly suited to growing oranges, while Chad has land in the desert that requires costly irrigation if it is to be used to grow oranges. Thus, Chad has a higher marginal cost of producing oranges.

The minimum acceptable prices that producers are willing to accept form points on a supply curve because the higher the price, the more bags of oranges will be supplied. At a price of \$3.50, for instance, only Carlos would be willing to supply a bag of oranges. But at a price of \$5.50, Carlos, Courtney, and Chuck would all be willing to supply a bag of oranges. The higher the market price, the more oranges will be supplied, as the market price surpasses the marginal costs and minimum acceptable prices of more and more producers. Thus, supply curves shown in this competitive market are both marginal-cost curves and minimum-acceptable-price curves.

The supply curve in Figure 4.2 includes not only the six producers named in Table 4.2 but also every other producer of oranges in the market. At the market price of \$8 per bag, Q_1 bags are produced because only those producers whose minimum acceptable prices are less than \$8 per bag will choose to produce oranges with their

FIGURE 4.2 **Producer surplus.** Producer surplus—shown as the blue triangle—is the difference between the actual price producers receive for a product (here \$8) and the lower minimum payments they are willing to accept. For quantity Q_1, producers receive the sum of the amounts represented by the blue triangle plus the yellow area. Because they need to receive only the amount shown by the yellow area to produce Q_1, the blue triangle represents producer surplus.



resources. Those lower acceptable prices for each of the units up to Q_1 are shown by the portion of the supply curve lying to the left of and below the assumed \$8 market price.

The individual producer surplus of each of these sellers is thus the vertical distance from each seller's respective minimum acceptable price on the supply curve up to the \$8 market price. Their collective producer surplus is shown by the blue triangle in Figure 4.2. In that figure, producers collect revenues of $P_1 \times Q_1$, which is the sum of the blue triangle and the yellow area. As shown by the supply curve, however, revenues of only those illustrated by the yellow area would be required to entice producers to offer Q_1 bags of oranges for sale. The sellers therefore receive a producer surplus shown by the blue triangle. That surplus is the sum of the vertical distances between the supply curve and the \$8 equilibrium price at each of the quantities to the left of Q_1.

There is a direct (positive) relationship between equilibrium price and the amount of producer surplus. Given the supply curve, lower prices reduce producer surplus; higher prices increase it. If you pencil in a lower equilibrium price than \$8, you will see that the producer surplus triangle gets smaller. The gaps between the minimum acceptable payments and the actual prices narrow when the price falls. If you pencil in an equilibrium price

WORKED PROBLEMS

W4.1
Consumer and producer surplus

above \$8, the size of the producer surplus triangle increases. The gaps between minimum acceptable payments and actual prices widen when the price increases.

Efficiency Revisited

In Figure 4.3 we bring together the demand and supply curves of Figures 4.1 and 4.2 to show the equilibrium price and quantity and the previously described regions of consumer and producer surplus. All markets that have downsloping demand curves and upsloping supply curves yield consumer and producer surplus.

Because we are assuming in Figure 4.3 that the demand curve reflects buyers' full willingness to pay and the supply curve reflects all of the costs facing sellers, the equilibrium quantity in Figure 4.3 reflects economic efficiency, which consists of productive efficiency and allocative efficiency.

- ***Productive efficiency*** is achieved because competition forces orange growers to use the best technologies and combinations of resources available. Doing so minimizes the per-unit cost of the output produced.
- ***Allocative efficiency*** is achieved because the correct quantity of oranges—Q_1—is produced relative to other goods and services.

There are two ways to understand why Q_1 is the correct quantity of oranges. Both involve realizing that any resources directed toward the production of oranges are resources that could have been used to produce other products. Thus, the only way to justify taking any amount of any resource (land, labor, capital, entrepreneurship) away from the production of other products is if it brings more utility or satisfaction when devoted to the production of oranges than it would if it were used to produce other products.

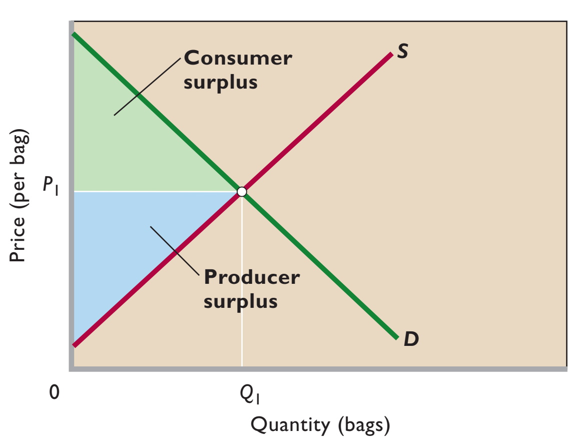
The first way to see why Q_1 is the allocatively efficient quantity of oranges is to note that demand and supply curves can be interpreted as measuring marginal benefit (MB) and marginal cost (MC). Recall from the discussion relating to Figure 1.3 that optimal allocation is achieved at the output level where MB = MC. We have already seen that supply curves are marginal cost curves. As it turns out, demand curves are marginal benefit curves. This is true because the maximum price that a consumer would be willing to pay for any particular unit is equal to the benefit that she would get if she were to consume that unit. Thus, each point on a demand curve represents both some consumer's maximum willingness to pay as well as the marginal benefit that he or she would get from consuming the particular unit in question.

Combining the fact that supply curves are MC curves with the fact that demand curves are MB curves, we see that points on the demand curve in Figure 4.3 measure the marginal benefit of oranges at each level of output, while points on the supply curve measure the marginal cost of oranges at each level of output. As a result, MB = MC where the demand and supply curves intersect—which means that the equilibrium quantity Q_1 must be allocatively efficient.

To gain a deeper understanding of why Q_1 is allocatively efficient, notice that for every unit up to Q_1 marginal benefit exceeds marginal cost (MB > MC). And because marginal cost includes the opportunity cost of not making other things with the resources needed to make these units, we know that people are made better off when the resources necessary to make these units are allocated to producing oranges rather than to producing anything else.

The second way to see why Q_1 is the correct quantity of oranges is based on our analysis of consumer and producer surplus and the fact that we can interpret demand and supply curves in terms of maximum willingness to pay and minimum acceptable price. In Figure 4.3, the maximum willingness to pay on the demand curve for each bag of oranges up to Q_1 exceeds the corresponding minimum acceptable price on the supply curve. Thus, each of these bags adds a positive amount (= maximum willingness to pay *minus* minimum acceptable price) to the *total* of consumer and producer surplus.

FIGURE 4.3 Efficiency: maximum combined consumer and producer surplus. At quantity Q_1 the combined amount of consumer surplus, shown as the green triangle, and producer surplus, shown as the blue triangle, is maximized. Efficiency occurs because, at Q_1, maximum willingness to pay, indicated by the points on the demand curve, equals minimum acceptable price, shown by the points on the supply curve.



The fact that maximum willingness to pay exceeds minimum acceptable price for every unit up to Q_1 means that people gain more utility from producing and consuming those units than they would if they produced and consumed anything else that could be made with the resources that went into making those units. This is true because both the maximum willingness to pay and the minimum acceptable price take opportunity costs into account. As long as the maximum willingness to pay exceeds the minimum acceptable price, people are willing to pay more to consume a unit of the good in question (here, bags of oranges) than they would pay to consume anything else that could be made with the same resources. Only at the equilibrium quantity Q_1—where the maximum willingness to pay exactly equals the minimum acceptable price—does society exhaust all opportunities to produce units for which benefits exceed costs (including opportunity costs). Producing Q_1 units therefore achieves allocative efficiency because the market is producing and distributing only those units that make people happier with bags of oranges than they would be with anything else that could be produced with the same resources.

Geometrically, producing Q_1 units maximizes the combined area of consumer and producer surplus in Figure 4.3. In this context, the combined area is referred to as *total surplus*. Thus, when Q_1 units are produced, total surplus is equal to the large triangle formed by the green consumer-surplus triangle and the blue producer-surplus triangle.

When demand curves reflect buyers' full willingness to pay and when supply curves reflect all the costs facing sellers, competitive markets produce equilibrium quantities that maximize the sum of consumer and producer surplus. Allocative efficiency occurs at the market equilibrium quantity where three conditions exist simultaneously:

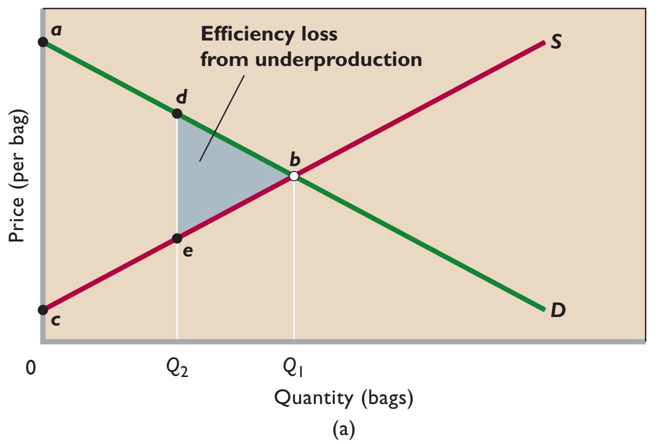
- MB = MC (Figure 1.3).
- Maximum willingness to pay = minimum acceptable price.
- Total surplus (= sum of consumer and producer surplus) is at a maximum.

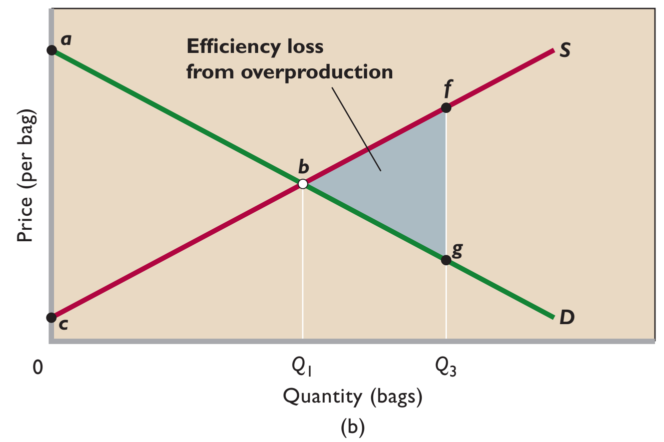
Economists are enamored of markets because properly functioning markets automatically achieve allocative efficiency. Other methods of allocating resources—such as government central planning—do exist. But because other methods cannot do any better than properly functioning markets—and may, in many cases, do much worse—economists usually prefer that resources be allocated through markets whenever properly functioning markets are available.

Efficiency Losses (or Deadweight Losses)

Figures 4.4a and 4.4b demonstrate that **efficiency losses**—reductions of combined consumer and producer surplus—result from both underproduction and overproduction. First, consider Figure 4.4a, which analyzes the case of underproduction by considering what happens if output falls from the efficient level Q_1 to the smaller amount Q_2. When that happens, the sum of consumer and producer surplus, previously *abc*, falls to *adec*. So the combined consumer and producer surplus declines by the amount of the gray triangle to the left of Q_1. That triangle represents an efficiency loss to buyers and sellers. And because buyers

FIGURE 4.4 Efficiency losses (or deadweight losses). Quantity levels either less than or greater than the efficient quantity Q_1 create efficiency losses. (a) Triangle *dbe* shows the efficiency loss associated with underproduction at output Q_2. (b) Triangle *bfg* illustrates the efficiency loss associated with overproduction at output level Q_3.





and sellers are members of society, it represents an efficiency loss (or a so-called **deadweight loss**) to society.

For output levels from Q_2 to Q_1, consumers' maximum willingness to pay (as reflected by points on the demand curve) exceeds producers' minimum acceptable price (as reflected by points on the supply curve). By failing to produce units of this product for which a consumer is willing to pay more than a producer is willing to accept, society suffers a loss of net benefits. As a concrete example, consider a particular unit for which a consumer is willing to pay $10 and a producer is willing to accept $6. The $4 difference between those values is a net benefit that will not be realized if this unit is not produced. In addition, the resources that should have gone to producing this unit will go instead to producing other products that will not generate as much utility as if those resources had been used here to produce this unit of this product. The triangle *dbe* in Figure 4.4a shows the total loss of net benefits that results from failing to produce the units from Q_2 to Q_1.

In contrast, consider the case of overproduction shown in Figure 4.4b, in which the number of oranges produced is Q_3 rather than the efficient level Q_1. In Figure 4.4b the combined consumer and producer surplus therefore declines by *bfg*—the gray triangle to the right of Q_1. This triangle subtracts from the total consumer and producer surplus of *abc* that would occur if the quantity had been Q_1. That is, for all units from 0 to Q_1, benefits exceed costs, so that those units generate the economic surplus shown by triangle *abc*. But the units from Q_1 to Q_3 are such that costs exceed benefits. Thus, they generate an economic loss shown by triangle *bfg*. The total economic surplus for all units from 0 to Q_3 is therefore the economic surplus given by *abc* for the units from 0 to Q_1 *minus* the economic loss given by *bfg* for the units from Q_1 to Q_3.

Producing any unit beyond Q_1 generates an economic loss because the willingness to pay for such units on the part of consumers is less than the minimum acceptable price to produce such units on the part of producers. As a concrete example, note that producing an item for which the maximum willingness to pay is, say, $7 and the minimum acceptable price is, say, $10 subtracts $3 from society's net benefits. Such production is uneconomical and creates an efficiency loss (or deadweight loss) for society. Because the net benefit of each bag of oranges from Q_1 to Q_3 is negative, we know that the benefits from these units are smaller than the opportunity costs of the other products that could have been produced with the resources that were used to produce these bags of oranges. The resources used to produce the bags from Q_1 to Q_3 could have generated net benefits instead of net losses if they had been directed toward producing other products. The gray triangle *bfg* to the right of Q_1 in Figure 4.4b shows the total efficiency loss from overproduction at Q_3.

The magic of markets is that when demand reflects consumers' full willingness to pay and when supply reflects all costs, the market equilibrium quantity will automatically equal the allocatively efficient output level. Under these conditions, the market equilibrium quantity will ensure that there are neither efficiency losses from underproduction nor efficiency losses from overproduction. As we are about to see, however, such losses do happen when either demand does not reflect consumers' full willingness to pay or supply does not reflect all costs.

QUICK REVIEW 4.1

- Market failures in competitive markets have two possible causes: demand curves that do not reflect consumers' full willingness to pay and supply curves that do not reflect producers' full cost of production.
- Consumer surplus is the difference between the maximum price that a consumer is willing to pay for a product and the lower price actually paid.
- Producer surplus is the difference between the minimum price that a producer is willing to accept for a product and the higher price actually received.
- At the equilibrium price and quantity in competitive markets, marginal benefit equals marginal cost, maximum willingness to pay equals minimum acceptable price, and the total of consumer surplus and producer surplus is maximized. Each of these conditions defines allocative efficiency.
- Quantities less than or greater than the allocatively efficient level of output create efficiency losses, often called deadweight losses.

Public Goods

LO4.3 Describe free riding and public goods, and illustrate why private firms cannot normally produce public goods.

Demand-side market failures arise in competitive markets when demand curves fail to reflect consumers' full willingness to pay for a good or service. In such situations, markets fail to produce all of the units for which there are net benefits because demand curves underreport how much consumers are willing and able to pay. This underreporting problem reaches its most extreme form in the case of a public good: Markets may fail to produce *any* of the public good because its demand curve may reflect *none* of its consumers' willingness to pay.

To understand public goods, we first need to understand the characteristics that define private goods.

Private Goods Characteristics

We have seen that the market system produces a wide range of **private goods.** These are the goods offered for sale in stores, in shops, and on the Internet. Examples include automobiles, clothing, personal computers, household appliances, and sporting goods. Private goods are distinguished by rivalry and excludability.

- **Rivalry** (in consumption) means that when one person buys and consumes a product, it is not available for another person to buy and consume. When Adams purchases and drinks a bottle of mineral water, it is not available for Benson to purchase and consume.
- **Excludability** means that sellers can keep people who do not pay for a product from obtaining its benefits. Only people who are willing and able to pay the market price for bottles of water can obtain these drinks and the benefits they confer.

Consumers fully express their personal demands for private goods in the market. If Adams likes bottled mineral water, that fact will be known by her desire to purchase the product. Other things equal, the higher the price of bottled water, the fewer bottles she will buy. So Adams's demand for bottled water will reflect an inverse relationship between the price of bottled water and the quantity of it demanded. This is simply *individual* demand, as described in Chapter 3.

The *market* demand for a private good is the horizontal summation of the individual demand schedules (review Figure 3.2). Suppose just two consumers comprise the market for bottled water and the price is \$1 per bottle. If Adams will purchase 3 bottles and Benson will buy 2, the market demand will reflect consumers' demand for 5 bottles at the \$1 price. Similar summations of quantities demanded at other prices will generate the market demand schedule and curve.

Suppose the equilibrium price of bottled water is \$1. Adams and Benson will buy a total of 5 bottles, and the sellers will obtain total revenue of \$5 (= \$1 × 5). If the sellers' cost per bottle is \$0.80, their total cost will be \$4 (= \$0.80 × 5). So sellers charging \$1 per bottle will obtain \$5 of total revenue, incur \$4 of total cost, and earn \$1 of profit on the 5 bottles sold.

Because firms can profitably "tap market demand" for private goods, they will produce and offer them for sale. Consumers demand private goods, and profit-seeking suppliers produce goods that satisfy the demand. Consumers willing to pay the market price obtain the goods; nonpayers go without. A competitive market not only makes private goods available to consumers but also allocates society's resources efficiently to the particular product. There is neither underproduction nor overproduction of the product.

Public Goods Characteristics

Public goods have the opposite characteristics of private goods. Public goods are distinguished by nonrivalry and nonexcludability.

- **Nonrivalry** (in consumption) means that one person's consumption of a good does not preclude consumption of the good by others. Everyone can simultaneously obtain the benefit from a public good such as national defense, street lighting, a global positioning system, or environmental protection.
- **Nonexcludability** means there is no effective way of excluding individuals from the benefit of the good once it comes into existence. Once in place, you cannot exclude someone from benefiting from national defense, street lighting, a global positioning system, or environmental protection.

These two characteristics create a **free-rider problem.** Once a producer has provided a public good, everyone, including nonpayers, can obtain the benefit.

Because most people do not voluntarily pay for something that they can obtain for free, most people become free riders. These free riders like the public good and would be willing to pay for it if producers could somehow force them to pay—but nonexcludability means that there is no way for producers to withhold the good from the free riders without also denying it to the few who do pay. As a result, free riding means that the willingness to pay of the free riders is not expressed in the market. From the viewpoint of producers, free riding reduces demand. The more free riding, the less demand. And if all consumers free ride, demand will collapse all the way to zero.

The low or even zero demand caused by free riding makes it virtually impossible for private firms to profitably provide public goods. With little or no demand, firms cannot effectively "tap market demand" for revenues and profits. As a result, they will not produce public goods. Society will therefore suffer efficiency losses because

CONSIDER THIS . . .

Street Entertainers

Street entertainers are often found in tourist areas of major cities. These entertainers illuminate the concepts of free riders and public goods.

Most street entertainers have a hard time earning a living from their activities (unless event organizers pay them) because they have no way of excluding nonpayers from the benefits of their entertainment. They essentially are providing public, not private, goods and must rely on voluntary payments.

The result is a significant free-rider problem. Only a few in the audience put money in the container or instrument case, and many who do so contribute only token amounts. The rest are free riders who obtain the benefits of the street entertainment and retain their money for purchases that they initiate.

Street entertainers are acutely aware of the free-rider problem, and some have found creative ways to lessen it. For example, some entertainers involve the audience directly in the act. This usually creates a greater sense of audience willingness (or obligation) to contribute money at the end of the performance.

"Pay for performance" is another creative approach to lessening the free-rider problem. A good example is the street entertainer painted up to look like a statue. When people drop coins into the container, the "statue" makes a slight movement. The greater the contributions, the greater the movement. But these human "statues" still face a free-rider problem: Nonpayers also get to enjoy the acts.

goods for which marginal benefits exceed marginal costs are not produced. Thus, if society wants a public good to be produced, it will have to direct government to provide it. Because the public good will still feature nonexcludability, the government won't have any better luck preventing free riding or charging people for it. But because the government can finance the provision of the public good through the taxation of other things, the government does not have to worry about profitability. It can therefore provide the public good even when private firms can't.

Examples of public goods include national defense, outdoor fireworks displays, the light beams thrown out by lighthouses, public art displays, public music concerts, MP3 music files posted to file-sharing Web sites, and ideas and inventions that are not protected by patents or copyrights. Each of these goods or services shows both nonrivalry and nonexcludability.

In a few special cases, private firms can provide public goods because the production costs of these public goods can be covered by the profits generated by closely related private goods. For instance, private companies can make a profit providing broadcast TV—which is a nonrival, nonexcludable public good—because they control who gets to air TV commercials, which are rival and excludable private goods. The money that broadcasters make from selling airtime for ads allows them to turn a profit despite having to give their main product, broadcast TV, away for free.

Unfortunately, only a few public goods can be subsidized in this way by closely related private goods. For the large majority of public goods, private provision is unprofitable. As a result, there are only two remaining ways for a public good to be provided: private philanthropy or government provision. For many less expensive or less important public goods like fireworks displays or public art, society may feel comfortable relying on private philanthropy. But when it comes to public goods like national defense, people normally look to the government.

This leads to an important question: Once a government decides to produce a particular public good, how can it determine the optimal amount that it should produce? How can it avoid either underallocating or overallocating society's scarce resources to the production of the public good?

Optimal Quantity of a Public Good

If consumers need not reveal their true demand for a public good in the marketplace, how can society determine the optimal amount of that good? The answer is that the government has to try to estimate the demand for a public good through surveys or public votes. It can then compare the marginal benefit (MB) of an added unit of the good against the government's marginal cost (MC) of providing it. Adhering to the MB = MC rule, government can provide the "right," meaning "efficient," amount of the public good.

Demand for Public Goods

The demand for a public good is somewhat unusual. Suppose Adams and Benson are the only two people in the society, and their marginal willingness to pay for a public good, national defense, is as shown in columns 1 and 2 and columns 1 and 3 in Table 4.3. Economists might have

TABLE 4.3 **Demand for a Public Good, Two Individuals**

(1) Quantity of Public Good	(2) Adams's Willingness to Pay (Price)		(3) Benson's Willingness to Pay (Price)		(4) Collective Willingness to Pay (Price)
1	$4	+	$5	=	$9
2	3	+	4	=	7
3	2	+	3	=	5
4	1	+	2	=	3
5	0	+	1	=	1

discovered these schedules through a survey asking hypothetical questions about how much each citizen was willing to pay for various types and amounts of public goods rather than go without them.

Notice that the schedules in Table 4.3 are price-quantity schedules, implying that they are demand schedules. Rather than depicting demand in the usual way—the quantity of a product someone is willing to buy at each possible price—these schedules show the price someone is willing to pay for an extra unit at each possible quantity. That is, Adams is willing to pay $4 for the first unit of the public good, $3 for the second, $2 for the third, and so on.

Suppose the government produces 1 unit of this public good. Because of nonrivalry, Adams's consumption of the good does not preclude Benson from also consuming it, and vice versa. So both consume the good, and neither volunteers to pay for it. But from Table 4.3 we can find the amount these two people would be willing to pay, together, rather than do without this 1 unit of the good. Columns 1 and 2 show that Adams would be willing to pay $4 for the first unit of the public good; columns 1 and 3 show that Benson would be willing to pay $5 for it. So the two people are jointly willing to pay $9 (= $4 + $5) for this first unit.

For the second unit of the public good, the collective price they are willing to pay is $7 (= $3 from Adams + $4 from Benson); for the third unit they would pay $5 (= $2 + $3); and so on. By finding the collective willingness to pay for each additional unit (column 4), we can construct a collective demand schedule (a willingness-to-pay schedule) for the public good. Here we are *not* adding the quantities demanded at each possible price, as we do when we determine the market demand for a private good. Instead, we are adding the prices that people are willing to pay for the last unit of the public good at each possible quantity demanded.

Figure 4.5 shows the same adding procedure graphically, using the data from Table 4.3. Note that we sum Adams's and Benson's willingness-to-pay curves *vertically* to derive the collective willingness-to-pay curve (demand curve). The summing procedure is downward from the top graph to the middle graph to the bottom (total) graph. For example, the height of the collective demand curve D_c at 2 units of output in the bottom graph is $7, the sum of the amounts that Adams and Benson are each willing to pay for the second unit (= $3 + $4). Likewise, the height of the collective demand curve at 4 units of the public good is $3 (= $1 + $2).

What does it mean in Figure 4.5a that, for example, Adams is willing to pay $3 for the second unit of the public good? It means that Adams expects to receive $3 of extra benefit or utility from that unit. And we know from our discussion of diminishing marginal utility in Chapter 3 that successive units of any good yield less and less added

CONSIDER THIS . . .

Responding to Digital Free Riding

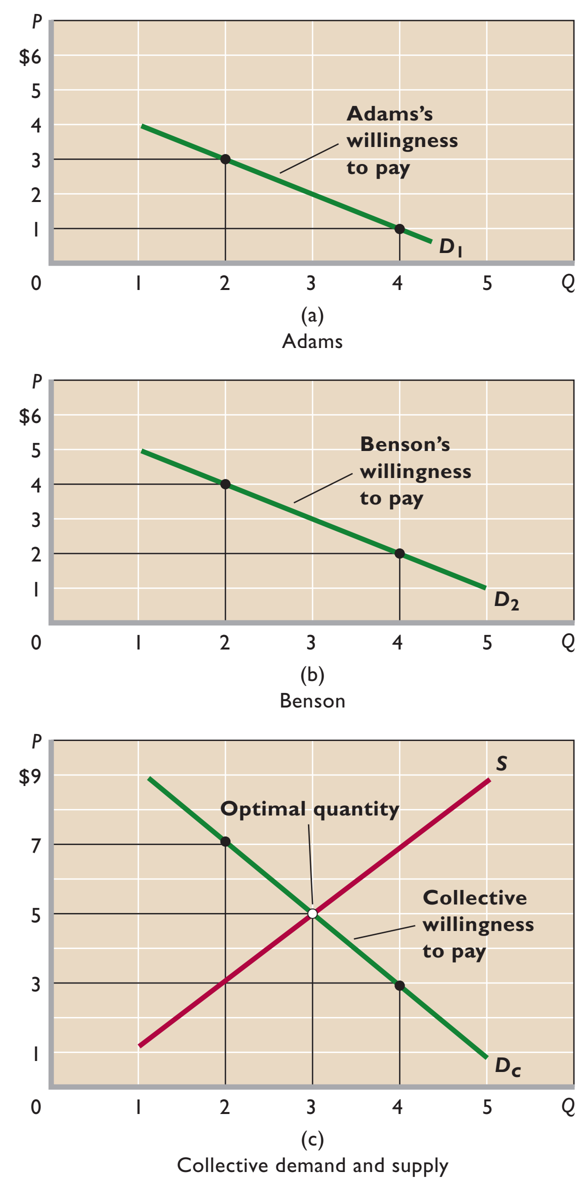
Four teenage friends start a rock band. They practice hard, master their instruments, write their own songs, and do gig after gig for nearly nothing at local bars to gain experience and perfect their music.

After nearly five years of effort, they get signed to a major record label. But the year is 2005 and record sales are collapsing due to digital piracy. The rise of Internet file sharing has turned music into a public good and sales of recorded music have collapsed as hundreds of millions of music lovers have become digital free riders.

At first, the band struggles with the new reality. If they can't make a living selling music, they might have to quit music and get regular jobs. But then they realize that while recorded music is now free for anyone who wants it to be free, live music isn't. And neither are T-shirts or memorabilia.

So the band promotes itself online and allows free downloads to help propel its popularity. But then it charges steep prices at live concerts and makes sure that its T-shirts and memorabilia also generate substantial revenues. By doing so, the band adjusts to the new reality in which music has become a public good, but live concerts and T-shirts have not. They charge for the items that are still private goods.

FIGURE 4.5 The optimal amount of a public good. Two people—Adams and Benson—are the only members of a hypothetical economy. (a) D_1 shows Adams's willingness to pay for various quantities of a particular public good. (b) D_2 shows Benson's willingness to pay pay for these same quantities of this public good. (c) The collective demand for this public good is shown by D_c and is found by summing vertically Adams's and Benson's individual willingness-to-pay curves. The supply (S) of the public good is upsloping, reflecting rising marginal costs. The optimal amount of the public good is 3 units, determined by the intersection of D_c and S. At that output, marginal benefit (reflected in the collective demand curve D_c) equals marginal cost (reflected in the supply curve S).



benefit. This is also true for public goods, explaining the downward slope of the willingness-to-pay curves of Adams, Benson, and society. These curves, in essence, are marginal-benefit (MB) curves.

Comparing MB and MC

We can now determine the optimal quantity of the public good. The collective demand curve D_c in Figure 4.5c measures society's marginal benefit of each unit of this particular good. The supply curve S in that figure measures society's marginal cost of each unit. The optimal quantity of this public good occurs where marginal benefit equals marginal cost, or where the two curves intersect. In Figure 4.5c that point is 3 units of the public good, where the collective willingness to pay for the last (third) unit—the marginal benefit—just matches that unit's marginal cost ($5 = $5). As we saw in Chapter 1, equating marginal benefit and marginal cost efficiently allocates society's scarce resources.

WORKED PROBLEMS

W4.2

Optimal amount of a public good

Cost-Benefit Analysis

The above example suggests a practical means, called **cost-benefit analysis,** for deciding whether to provide a particular public good and how much of it to provide. Like our example, cost-benefit analysis (or marginal-benefit–marginal-cost analysis) involves a comparison of marginal costs and marginal benefits.

Concept Suppose the federal government is contemplating a highway construction plan. Because the economy's resources are limited, any decision to use more resources in the public sector will mean fewer resources for the private sector. There will be an opportunity cost, as well as a benefit. The cost is the loss of satisfaction resulting from the accompanying decline in the production of private goods; the benefit is the extra satisfaction resulting from the output of more public goods. Should the needed resources be shifted from the private to the public sector? The answer is yes if the benefit from the extra public goods exceeds the cost that results from having fewer private goods. The answer is no if the cost of the forgone private goods is greater than the benefit associated with the extra public goods.

Cost-benefit analysis, however, can indicate more than whether a public program is worth doing. It can also help the government decide on the *extent* to which a project should be pursued. Real economic questions cannot usually be answered simply by "yes" or "no" but, rather, involve questions such as "how much" or "how little."

Illustration Roads and highways can be run privately, as excludability is possible with toll gates. However, the federal highway system is almost entirely nonexclusive

TABLE 4.4 **Cost-Benefit Analysis for a National Highway Construction Project (in Billions)**

(1) Plan	(2) Total Cost of Project	(3) Marginal Cost	(4) Total Benefit	(5) Marginal Benefit	(6) Net Benefit (4) − (2)
No new construction	$ 0		$ 0		$ 0
		$ 4		$ 5	
A: Widen existing highways	4		5		1
		6		8	
B: New 2-lane highways	10		13		3
		8		**10**	
C: New 4-lane highways	**18**		**23**		**5**
		10		3	
D: New 6-lane highways	28		26		−2

because anyone with a car can get on and off most federal highways without restriction anytime they want. Federal highways therefore satisfy one characteristic of a public good, nonexcludability. The other characteristic, nonrivalry, is also satisfied by the fact that unless a highway is already extremely crowded, one person's driving on the highway does not preclude another person's driving on the highway. Thus, the federal highway system is effectively a public good. This leads us to ask: Should the federal government expand the federal highway system? If so, what is the proper size or scope for the overall project?

Table 4.4 lists a series of increasingly ambitious and increasingly costly highway projects: widening existing two-lane highways; building new two-lane highways; building new four-lane highways; building new six-lane highways. The extent to which government should undertake highway construction depends on the costs and benefits. The costs are largely the costs of constructing and maintaining the highways; the benefits are improved flows of people and goods throughout the country.[2]

The table shows that total annual benefit (column 4) exceeds total annual cost (column 2) for plans A, B, and C, indicating that some highway construction is economically justifiable. We see this directly in column 6, where total costs (column 2) are subtracted from total annual benefits (column 4). Net benefits are positive for plans A, B, and C. Plan D is not economically justifiable because net benefits are negative.

But the question of optimal size or scope for this project remains. Comparing the marginal cost (the change in total cost) and the marginal benefit (the change in total benefit) relating to each plan determines the answer. The guideline is well known to you from previous discussions: Increase an activity, project, or output as long as the marginal benefit (column 5) exceeds the marginal cost (column 3). Stop the activity at, or as close as possible to, the point at which the marginal benefit equals the marginal cost. Do not undertake a project for which marginal cost exceeds marginal benefit.

In this case plan C (building new four-lane highways) is the best plan. Plans A and B are too modest; the marginal benefits exceed the marginal costs, and there is a better option. Plan D's marginal cost ($10 billion) exceeds the marginal benefit ($3 billion) and therefore cannot be justified; it overallocates resources to the project. Plan C is closest to the theoretical optimum because its marginal benefit ($10 billion) still exceeds marginal cost ($8 billion) but approaches the MB = MC (or MC = MB) ideal.

This **marginal-cost–marginal-benefit rule** actually tells us which plan provides the maximum excess of total benefits over total costs or, in other words, the plan that provides society with the maximum net benefit. You can confirm directly in column 6 that the maximum net benefit (= $5 billion) is associated with plan C.

Cost-benefit analysis shatters the myth that "economy in government" and "reduced government spending" are synonymous. "Economy" is concerned with using scarce resources efficiently. If the marginal cost of a proposed government program exceeds its marginal benefit, then the proposed public program should not be undertaken. But if the marginal benefit exceeds the marginal cost, then it would be uneconomical or "wasteful" not to spend on that government program. Economy in government does not mean minimization of public spending. It means allocating resources between the private and public sectors and among public goods to achieve maximum net benefit.

Quasi-Public Goods

Government provides many goods that fit the economist's definition of a public good. However, it also provides other goods and services that could be produced and delivered in such a way that exclusion would be possible. Such goods, called **quasi-public goods,** include education, streets and highways, police and fire protection, libraries and museums, preventive medicine, and sewage disposal. They could all be

[2]Because the costs of public goods typically are immediate while the benefits often accrue over longer time periods, economists convert both costs and benefits to present values for comparison. Doing so properly accounts for the time-value of money, discussed at length in later chapters.

priced and provided by private firms through the market system. But, because the benefits of these goods flow well beyond the benefit to individual buyers, these goods would be underproduced by the market system. Therefore, government often provides them to avoid the underallocation of resources that would otherwise occur.

The Reallocation Process

How are resources reallocated from the production of private goods to the production of public and quasi-public goods? If the resources of the economy are fully employed, government must free up resources from the production of private goods and make them available for producing public and quasi-public goods. It does so by reducing private demand for them. And it does that by levying taxes on households and businesses, taking some of their income out of the circular flow. With lower incomes and hence less purchasing power, households and businesses must curtail their consumption and investment spending. As a result, the private demand for goods and services declines, as does the private demand for resources. So by diverting purchasing power from private spenders to government, taxes remove resources from private use.

Government then spends the tax proceeds to provide public and quasi-public goods and services. Taxation releases resources from the production of private consumer goods (food, clothing, television sets) and private investment goods (printing presses, boxcars, warehouses). Government shifts those resources to the production of public and quasi-public goods (post offices, submarines, parks), changing the composition of the economy's total output.

QUICK REVIEW 4.2

- Public goods are characterized by nonrivalry and nonexcludability.
- The demand (marginal-benefit) curve for a public good is found by vertically adding the prices that all the members of society are willing to pay for the last unit of output at various output levels.
- The socially optimal amount of a public good is the amount at which the marginal cost and marginal benefit of the good are equal.
- Cost-benefit analysis is the method of evaluating alternative projects or sizes of projects by comparing the marginal cost and marginal benefit and applying the MC = MB rule.
- The government uses taxes to reallocate resources from the production of private goods to the production of public and quasi-public goods.

Externalities

LO4.4 Explain how positive and negative externalities cause under- and overallocations of resources.

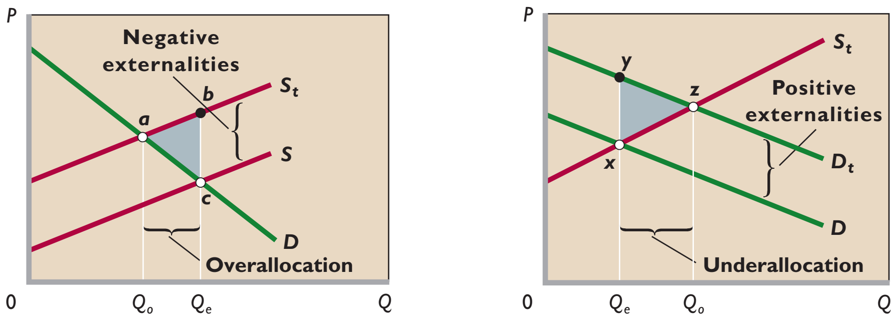
In addition to providing public goods, governments can also improve the allocation of resources in the economy by correcting for market failures caused by externalities. An **externality** occurs when some of the costs or the benefits of a good or service are passed onto or "spill over to" someone other than the immediate buyer or seller. Such spillovers are called externalities because they are benefits or costs that accrue to some third party that is external to the market transaction.

There are both positive and negative externalities. An example of a negative externality is the cost of breathing polluted air; an example of a positive externality is the benefit of having everyone else inoculated against some disease. When there are negative externalities, an overproduction of the related product occurs and there is an overallocation of resources to this product. Conversely, underproduction and underallocation of resources result when positive externalities are present.

Negative Externalities

Negative externalities cause supply-side market failures. These failures happen because producers do not take into account the costs that their negative externalities impose on others. This failure to account for all production costs causes firms' supply curves to shift to the right of (or below) where they would be if firms properly accounted for all costs. Consider the costs of breathing polluted air that are imposed on third parties living downwind of smoke-spewing factories. Because polluting firms do not take account of such costs, they oversupply the products they make, producing units for which total costs (including those that fall on third parties) exceed total benefits. The same is true when airlines fail to account for the costs that noisy jet engines impose on people living near airports and when biodiesel factories that convert dead animal parts into fuel release foul smelling gases that disgust those living nearby.

Figure 4.6a illustrates how negative externalities affect the allocation of resources. When producers shift some of their costs onto the community as external costs, producers' marginal costs are lower than they would be if they had to pay for those costs. So their supply curves do not include or "capture" all the costs legitimately associated with the production of their goods. A polluting producer's supply curve such as S in Figure 4.6a therefore understates the total cost of production. The firm's supply curve lies to the right of (or below) the total-cost supply curve S_t, which



(a)
Negative externalities

(b)
Positive externalities

FIGURE 4.6 Negative externalities and positive externalities. (a) With negative externalities borne by society, the producers' supply curve *S* is to the right of (below) the total-cost supply curve S_t. Consequently, the equilibrium output Q_e is greater than the optimal output Q_o, and the efficiency loss is *abc*. (b) When positive externalities accrue to society, the market demand curve *D* is to the left of (below) the total-benefit demand curve D_t. As a result, the equilibrium output Q_e is less than the optimal output Q_o, and the efficiency loss is *xyz*.

would include the spillover cost. Through polluting and thus transferring costs to society, the firm enjoys lower production costs and has the supply curve *S*.

The outcome is shown in Figure 4.6a, where equilibrium output Q_e is larger than the optimal output Q_o. This means that resources are overallocated to the production of this commodity; too many units of it are produced. In fact, there is a net loss to society for every unit from Q_o to Q_e because, for those units, the supply curve that accounts for all costs, S_t, lies above the demand curve. Therefore, MC exceeds MB for those units. The resources that went into producing those units should have been used elsewhere in the economy to produce other things.

In terms of our previous analysis, the negative externality results in an efficiency loss represented by triangle *abc*.

Positive Externalities

Positive externalities cause demand-side market failures. These failures happen because market demand curves in such cases fail to include the willingness to pay of the third parties who receive the external benefits caused by the positive externality. This failure to account for all benefits shifts market demand curves to the left of (or below) where they would be if they included all benefits and the willingness to pay of both the third parties as well as the primary beneficiaries. Because demand curves fail to take into account all benefits when there are positive externalities, markets in such cases fail to produce all units for which benefits (including those that are received by third parties) exceed costs. As a result, products featuring positive externalities are underproduced.

Vaccinations are a good example of how positive externalities reduce demand and shift demand curves down and to the left. When John gets vaccinated against a disease, he benefits not only himself (because he can no longer contract the disease) but also everyone else around him (because they know that in the future he will never be able to infect them). These other people would presumably be willing to pay some positive amount of money for the benefits they receive when John is vaccinated. But because his vaccination is a public good, there is no way to make them pay.

To see why his vaccination is a public good, note that the vaccination benefits that John provides to others feature nonrivalry and nonexcludability. There is nonrivalry because the protection his vaccination provides to one person does not lessen the protection that it provides to other people. There is nonexcludability because once he is vaccinated, there is no way to exclude anyone in particular from benefiting from his vaccination. Thus, the market demand for vaccinations will only include John's personal willingness to pay for the benefits that he personally receives from the vaccination. The market demand will fail to include the benefits that others receive. As a result, demand will be too low and vaccinations will be underproduced.

ORIGIN OF THE IDEA

O4.2

Externalities

Figure 4.6b shows the impact of positive externalities on resource allocation. When external benefits occur, the market demand curve *D* lies to the left of (or below) the total-benefits demand curve, D_t. That is, *D* does not include the external benefits of the product, whereas D_t does.

The outcome is that the equilibrium output Q_e is less than the optimal output Q_o. The market fails to produce enough vaccinations, and resources are underallocated to this product. The underproduction implies that society is missing out on a significant amount of potential net benefits. For every unit from Q_e to Q_o, the demand curve that accounts for all benefits, D_t, lies above the supply curve that accounts for all costs—including the opportunity cost of producing other items with the resources that would be needed to produce these units. Therefore, MB exceeds MC for each of these units, and we know that society should redeploy some of its resources away from the production of other things in order to produce these units that generate net benefits.

In terms of our previous analysis, the positive externality results in an efficiency loss represented by triangle *xyz*.

Government Intervention

Government intervention may be called upon to achieve economic efficiency when externalities affect large numbers of people or when community interests are at stake. Government can use direct controls and taxes to counter negative externalities; it may provide subsidies or public goods to deal with positive externalities.

Direct Controls The direct way to reduce negative externalities from a certain activity is to pass legislation limiting that activity. Such direct controls force the offending firms to incur the actual costs of the offending activity. Historically, direct controls in the form of uniform emission standards—limits on allowable pollution—have dominated American air pollution policy. For example, the Clean Air Act of 1990 (1) forced factories and businesses to install "maximum achievable control technology" to reduce emissions of 189 toxic chemicals by 90 percent between 1990 and 2000; (2) required a 30 to 60 percent reduction in tailpipe emissions from automobiles by 2000; (3) mandated a 50 percent reduction in the use of chlorofluorocarbons (CFCs), which deplete the ozone layer (CFCs were used widely as a coolant in refrigeration, a blowing agent for foam, and a solvent in the electronics industry); and (4) forced coal-burning utilities to cut their emissions of sulfur dioxide by about 50 percent to reduce the acid-rain destruction of lakes and forests. Clean-water legislation limits the amount of heavy metals, detergents, and other pollutants firms can discharge into rivers and bays. Toxic-waste laws dictate special procedures and dump sites for disposing of contaminated soil and solvents. Violating these laws means fines and, in some cases, imprisonment.

Direct controls raise the marginal cost of production because the firms must operate and maintain pollution-control equipment. The supply curve S in Figure 4.7b, which does not reflect the external costs, shifts leftward to the total-cost supply curve, S_t. Product price increases, equilibrium output falls from Q_e to Q_o, and the initial overallocation of resources shown in Figure 4.7a is corrected. Observe that the efficiency loss shown by triangle *abc* in Figure 4.7a disappears after the overallocation is corrected in Figure 4.7b.

CONSIDER THIS . . .

The Fable of the Bees

Economist Ronald Coase received the Nobel Prize for his so-called **Coase theorem,** which pointed out that under the right conditions, private individuals could often negotiate their own mutually agreeable solutions to externality problems through *private bargaining* without the need for government interventions like pollution taxes.

This is a very important insight because it means that we shouldn't automatically call for government intervention every time we see a potential externality problem. Consider the positive externalities that bees provide by pollinating farmers' crops. Should we assume that beekeeping will be underprovided unless the government intervenes with, for instance, subsidies to encourage more hives and hence more pollination?

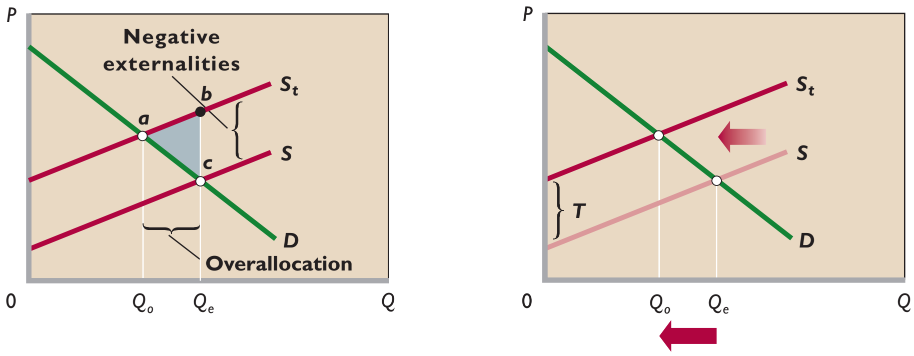
As it turns out, no. Research has shown that farmers and beekeepers long ago used private bargaining to develop customs and payment systems that avoid free riding by farmers and encourage beekeepers to keep the optimal number of hives. Free riding is avoided by the custom that all farmers in an area simultaneously hire beekeepers to provide bees to pollinate their crops. And farmers always pay the beekeepers for their pollination services because if they didn't, then no beekeeper would ever work with them in the future—a situation that would lead to massively reduced crop yields due to a lack of pollination.

ORIGIN OF THE IDEA

O4.3

Coase theorem

The "Fable of the Bees" is a good reminder that it is a fallacy to assume that the government must always get involved to remedy externalities. In many cases, the private sector can solve both positive and negative externality problems on its own.



(a)
Negative externalities

(b)
Correcting the overallocation of resources via direct controls or via a tax

FIGURE 4.7 Correcting for negative externalities. (a) Negative externalities result in an overallocation of resources. (b) Government can correct this overallocation in two ways: (1) using direct controls, which would shift the supply curve from S to S_t and reduce output from Q_e to Q_o, or (2) imposing a specific tax T, which would also shift the supply curve from S to S_t, eliminating the overallocation of resources and thus the efficiency loss.

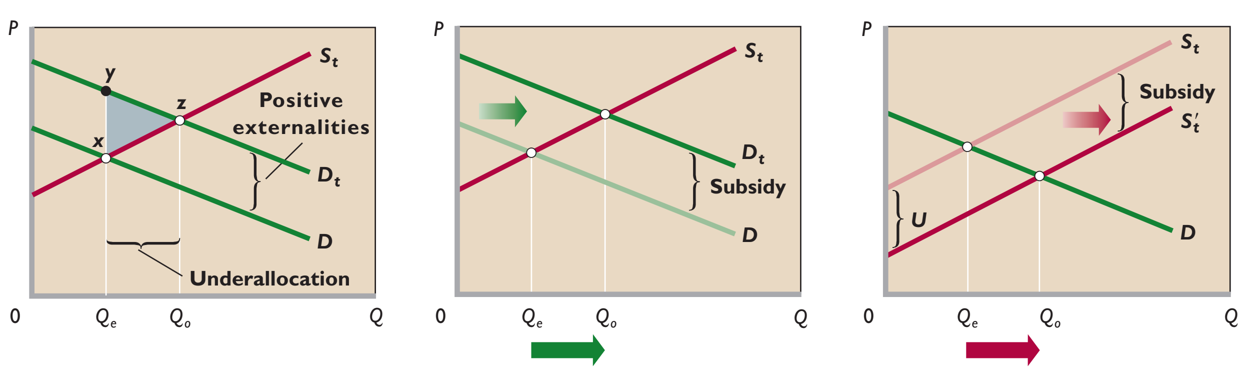
Specific Taxes A second policy approach to negative externalities is for government to levy taxes or charges specifically on the related good. For example, the government has placed a manufacturing excise tax on CFCs, which deplete the stratospheric ozone layer protecting the earth from excessive solar ultraviolet radiation. Facing such an excise tax, manufacturers must decide whether to pay the tax or expend additional funds to purchase or develop substitute products. In either case, the tax raises the marginal cost of producing CFCs, shifting the private supply curve for this product leftward (or upward).

In Figure 4.7b, a tax equal to T per unit increases the firm's marginal cost, shifting the supply curve from S to S_t. The equilibrium price rises, and the equilibrium output declines from Q_e to the economically efficient level Q_o. The tax thus eliminates the initial overallocation of resources and therefore the efficiency loss.

Subsidies and Government Provision Where spillover benefits are large and diffuse, as in our earlier example of inoculations, government has three options for correcting the underallocation of resources:

- ***Subsidies to buyers*** Figure 4.8a again shows the supply-demand situation for positive externalities. Government could correct the underallocation of

FIGURE 4.8 Correcting for positive externalities. (a) Positive externalities result in an underallocation of resources. (b) This underallocation can be corrected through a subsidy to consumers, which shifts market demand from D to D_t and increases output from Q_e to Q_o. (c) Alternatively, the underallocation can be eliminated by providing producers with a subsidy of U, which shifts their supply curve from S_t to $S_{t'}$, increasing output from Q_e to Q_o and eliminating the underallocation, and thus the efficiency loss, shown in graph a.



(a)
Positive externalities

(b)
Correcting the underallocation of resources via a subsidy to consumers

(c)
Correcting the underallocation of resources via a subsidy to producers

resources, for example, to inoculations, by subsidizing consumers of the product. It could give each new mother in the United States a discount coupon to be used to obtain a series of inoculations for her child. The coupon would reduce the "price" to the mother by, say, 50 percent. As shown in Figure 4.8b, this program would shift the demand curve for inoculations from too-low D to the appropriate D_t. The number of inoculations would rise from Q_e to the economically optimal Q_o, eliminating the underallocation of resources and efficiency loss shown in Figure 4.8a.

- ***Subsidies to producers*** A subsidy to producers is a tax in reverse. Taxes are payments *to* the government that increase producers' costs. Subsidies are payments *from* the government that decrease producers' costs. As shown in Figure 4.8c, a subsidy of U per inoculation to physicians and medical clinics would reduce their marginal costs and shift their supply curve rightward from S_t to S_t'. The output of inoculations would increase from Q_e to the optimal level Q_o, correcting the underallocation of resources and efficiency loss shown in Figure 4.8a.
- ***Government provision*** Finally, where positive externalities are extremely large, the government may decide to provide the product for free to everyone. The U.S. government largely eradicated the crippling disease polio by administering free vaccines to all children. India ended smallpox by paying people in rural areas to come to public clinics to have their children vaccinated.

Table 4.5 lists several methods for correcting externalities, including those we have discussed thus far.

Society's Optimal Amount of Externality Reduction

LO4.5 Show why we normally won't want to pay what it would cost to eliminate every last bit of a negative externality such as air pollution.

Negative externalities such as pollution reduce the utility of those affected, rather than increase it. These spillovers are not economic goods but economic "bads." If something is bad, shouldn't society eliminate it? Why should society allow firms or municipalities to discharge *any* impure waste into public waterways or to emit *any* pollution into the air?

Economists answer these questions by pointing out that reducing pollution and negative externalities is not free. There are costs as well as benefits to reducing pollution. As a result, the correct question to ask when it comes to cleaning up negative externalities is not, "Do we pollute a lot or pollute zero?" That is an all-or-nothing question that ignores marginal costs and marginal benefits. Instead, the correct question is, "What is the optimal amount to clean up—the amount that equalizes the marginal cost of cleaning up with the marginal benefit of a cleaner environment?"

If we ask that question, we see that reducing a negative externality has a "price." Society must decide how much of a reduction it wants to "buy." High costs may mean that totally eliminating pollution might not be desirable, even if it is technologically feasible. Because of the law of diminishing returns, cleaning up the second 10 percent of pollutants from an industrial smokestack normally is more costly than cleaning up the first 10 percent. Eliminating the third 10 percent is more costly than cleaning up the second 10 percent, and so on. Therefore, cleaning up the last 10 percent of pollutants is the most costly reduction of all.

TABLE 4.5 **Methods for Dealing with Externalities**

Problem	Resource Allocation Outcome	Ways to Correct
Negative externalities (spillover costs)	Overproduction of output and therefore overallocation of resources	1. Private bargaining 2. Liability rules and lawsuits 3. Tax on producers 4. Direct controls 5. Market for externality rights
Positive externalities (spillover benefits)	Underproduction of output and therefore underallocation of resources	1. Private bargaining 2. Subsidy to consumers 3. Subsidy to producers 4. Government provision

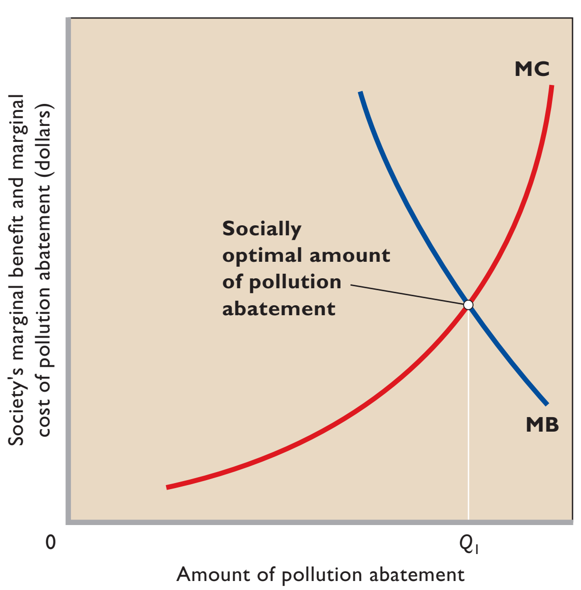
The marginal cost (MC) to the firm and hence to society—the opportunity cost of the extra resources used—rises as pollution is reduced more and more. At some point MC may rise so high that it exceeds society's marginal benefit (MB) of further pollution abatement (reduction). Additional actions to reduce pollution will therefore lower society's well-being; total cost will rise more than total benefit.

MC, MB, and Equilibrium Quantity

Figure 4.9 shows both the rising marginal-cost curve, MC, for pollution reduction and the downsloping marginal-benefit curve, MB, for pollution reduction. MB slopes downward because of the law of diminishing marginal utility: The more pollution reduction society accomplishes, the lower the utility (and benefit) of the next unit of pollution reduction.

The **optimal reduction of an externality** occurs when society's marginal cost and marginal benefit of reducing that externality are equal (MC = MB). In Figure 4.9 this optimal amount of pollution abatement is Q_1 units. When MB exceeds MC, additional abatement moves society toward economic efficiency; the added benefit of cleaner air or water exceeds the benefit of any alternative use of the required resources. When MC exceeds MB, additional abatement reduces economic efficiency; there would be greater benefits from using resources in some other way than to further reduce pollution.

FIGURE 4.9 Society's optimal amount of pollution abatement. The optimal amount of externality reduction—in this case, pollution abatement—occurs at Q_1, where society's marginal cost MC and marginal benefit MB of reducing the spillover are equal.



In reality, it is difficult to measure the marginal costs and benefits of pollution control. Nevertheless, Figure 4.9 demonstrates that some pollution may be economically efficient. This is so not because pollution is desirable but because beyond some level of control, further abatement may reduce society's net well-being. As an example, it would cost the government billions of dollars to clean up every last piece of litter in America. Thus, it would be better to tolerate some trash blowing around if the money saved by picking up less trash would yield larger net benefits when spent on other things.

Shifts in Locations of the Curves

The locations of the marginal-cost and marginal-benefit curves in Figure 4.9 are not forever fixed. They can, and probably do, shift over time. For example, suppose that the technology of pollution-control equipment improved noticeably. We would expect the cost of pollution abatement to fall, society's MC curve to shift rightward, and the optimal level of abatement to rise. Or suppose that society were to decide that it wanted cleaner air and water because of new information about the adverse health effects of pollution. The MB curve in Figure 4.9 would shift rightward, and the optimal level of pollution control would increase beyond Q_1. Test your understanding of these statements by drawing the new MC and MB curves in Figure 4.9.

Government's Role in the Economy

Market failures can be used to justify government interventions in the economy. The inability of private-sector firms to break even when attempting to provide public goods and the over- and underproduction problems caused by positive and negative externalities mean that government can have an important role to play if society's resources are to be efficiently allocated to the goods and services that people most highly desire.

Correcting for market failures is not, however, an easy task. To begin with, government officials must correctly identify the existence and the cause of any given market failure. That by itself may be difficult, time-consuming, and costly. But even if a market failure is correctly identified and diagnosed, government may still fail to take appropriate corrective action due to the fact that government undertakes its economic role in the context of politics.

To serve the public, politicians need to get elected. To stay elected, officials (presidents, senators, representatives, mayors, council members, school board members) need to satisfy their particular constituencies. At best, the political

LAST WORD

Carbon Dioxide Emissions, Cap and Trade, and Carbon Taxes

Cap-and-trade systems and carbon taxes are two approaches to reducing carbon dioxide (CO_2) emissions.

Externality problems are property rights problems. Consider a landfill. Because the owner of the landfill has full rights to his land, people wishing to dump their trash into the landfill have to pay him. This payment implies that there is no externality: He happily accepts their trash in exchange for a dumping fee. By contrast, because nobody owns the atmosphere, all air pollution is an externality, since there is no way for those doing the polluting to work out a payment to compensate those affected by the pollution or for those threatened with pollution to simply refuse to be polluted on.

Conventional property rights therefore cannot fix the externalities associated with air pollution. But that does not mean property rights can't help fight pollution. The trick to making them work is to assign property rights not to the atmosphere itself, but to *polluting* the atmosphere. This is done in "cap-and-trade" systems, under which the government sets an annual limit, or cap, to the number of tons of a pollutant that firms can emit into the atmosphere.

Consider carbon dioxide, or CO_2. It is a colorless, odorless gas that many scientists consider to be a contributing cause of climate change, specifically global warming. To reduce CO_2 emissions, the U.S. government might set a cap of 5 billion tons of CO_2 emissions per year in the United States (which would be about 25 percent below 2010 emissions levels for that molecule). The government then prints out emissions permits that sum to the limit set in the cap and distributes them to polluting firms. Once they are distributed, the only way a firm can legally emit a ton of CO_2 is if it owns a permit to do so.

Under this policy, the government can obviously adjust the total amount of air pollution by adjusting the cap. This by itself improves efficiency because the cap imposes scarcity. Because each firm has only a limited number of permits, each firm has a strong incentive to maximize the net benefit that it produces from every ton of pollution that it emits. But the *cap-and-trade* scheme leads to even greater improvements in efficiency because firms are free to trade (sell) them to each other in what are referred to as *markets for externality rights.*

For instance, suppose Smokestack Toys owns permits for 100 tons of CO_2 emissions and that it could use them to produce toy cars that would generate profits of $100,000. There is a power plant, however, that could make up to $1 million of profits by using those 100 tons of emissions permits to generate electricity. Because firms can trade their permits, Smokestack Toys will sell its permits to the power plant for more than the $100,000 in profits that it could make if it kept them and

realities complicate government's role in the economy; at worst, they produce undesirable economic outcomes.

In the political context, overregulation can occur in some cases; underregulation, in others. Some public goods and quasi-public goods can be produced not because their benefits exceed their costs but because their benefits accrue to firms located in states served by powerful elected officials. Inefficiency can easily creep into government activities because of the lack of a profit incentive to hold down costs. Policies to correct negative externalities can be politically blocked by the very parties that are producing the spillovers. In short, the economic role of government, although critical to a well-functioning economy, is not always perfectly carried out.

Economists use the term "government failure" to describe economically inefficient outcomes caused by shortcomings in the public sector.

QUICK REVIEW 4.3

- Policies for coping with the overallocation of resources, and therefore efficiency losses, caused by negative externalities are (a) private bargaining, (b) liability rules and lawsuits, (c) direct controls, (d) specific taxes, and (e) markets for externality rights (Last Word).
- Policies for correcting the underallocation of resources, and therefore efficiency losses, associated with positive externalities are (a) private bargaining, (b) subsidies to producers, (c) subsidies to consumers, and (d) government provision.
- The optimal amount of negative-externality reduction occurs where society's marginal cost and marginal benefit of reducing the externality are equal.
- Political pressures often lead government to respond inefficiently when attempting to correct for market failures.

produced toy cars. And the power plant will gladly pay more than $100,000 for those permits because it can turn around and use them to make up to $1 million of profits by using them to generate electricity.

Society will benefit hugely from this transaction because while 100 tons of CO_2 will be emitted no matter which firm uses the permits, society will receive much greater net benefits when they are used by the power plant, as indicated by the fact that the power plant can produce much larger profits than the toy company when using the same amount of this scarce resource.

Several words of caution, however! Cap-and-trade systems have proven very difficult to implement in cases where it is difficult for regulators to effectively check whether firms are obeying the system. This has been a major problem with the European Union's cap-and-trade system for CO_2 emissions. Because nearly every type of industrial activity releases CO_2 into the atmosphere, enforcement involves monitoring many thousands of factories of all sizes. That is very difficult and cheating has resulted. In addition, politically connected industries got politicians to give them exemptions or free permits.

By contrast, a cap-and-trade system on sulfur dioxide emissions from coal-burning public utilities has worked well in the United States since the 1980s. But in that case, there were only a few hundred polluting utilities, and they were already being monitored for emissions. So there was little ability to cheat. In addition, all of the firms were treated equally, with no firms allowed exemptions or free permits.

Due to the mixed results, many economists have concluded that a cap-and-trade system would not be the best way to curb CO_2 emissions in the United States. They believe that there are simply too many sources of pollution to make monitoring either possible or cost-effective. And it seems likely that politically connected industries will be granted exemptions. So, instead, many economists favor a carbon tax, which would involve taxing each ton of coal, each gallon of gasoline, and each barrel of oil on the basis of how much carbon it contains (and thus how much CO_2 will eventually be released into the atmosphere when it is used). By raising the cost of polluting, the tax would reduce consumption and lessen the externalities associated with CO_2 emissions. It would also be nearly impossible to evade, so that we would not have to worry about cheating.

SUMMARY

LO4.1 Differentiate between demand-side market failures and supply-side market failures.

A market failure happens in a particular market when the market produces an equilibrium level of output that either overallocates or underallocates resources to the product being traded in the market. In competitive markets that feature many buyers and many sellers, market failures can be divided into two types: Demand-side market failures occur when demand curves do not reflect consumers' full willingness to pay; supply-side market failures occur when supply curves do not reflect all production costs, including those that may be borne by third parties.

LO4.2 Explain the origin of both consumer surplus and producer surplus, and explain how properly functioning markets maximize their sum, total surplus, while optimally allocating resources.

Consumer surplus is the difference between the maximum price that a consumer is willing to pay for a product and the lower price actually paid; producer surplus is the difference between the minimum price that a producer is willing to accept for a product and the higher price actually received. Collectively, consumer surplus is represented by the triangle under the demand curve and above the actual price, whereas producer surplus is shown by the triangle above the supply curve and below the actual price.

Graphically, the combined amount of producer and consumer surplus is represented by the triangle to the left of the intersection of the supply and demand curves that is below the demand curve and above the supply curve. At the equilibrium price and quantity in competitive markets, marginal benefit equals marginal cost, maximum willingness to pay equals minimum acceptable price, and the combined amount of consumer surplus and producer surplus is maximized.

Output levels that are either less than or greater than the equilibrium output create efficiency losses, also called deadweight losses. These losses are reductions in the combined amount of consumer surplus and producer surplus. Underproduction creates efficiency losses because output is not being produced for which maximum willingness to pay exceeds minimum acceptable price. Overproduction creates efficiency losses because output is being produced for which minimum acceptable price exceeds maximum willingness to pay.

LO4.3 Describe free riding and public goods, and illustrate why private firms cannot normally produce public goods.

Public goods are distinguished from private goods. Private goods are characterized by rivalry (in consumption) and excludability. One person's purchase and consumption of a private good precludes others from also buying and consuming it. Producers can exclude nonpayers (free riders) from receiving the benefits. In contrast, public goods are characterized by nonrivalry (in consumption) and nonexcludability. Public goods are not profitable to private firms because nonpayers (free riders) can obtain and consume those goods without paying. Government can, however, provide desirable public goods, financing them through taxation.

The collective demand schedule for a particular public good is found by summing the prices that each individual is willing to pay for an additional unit. Graphically, that demand curve is found by summing vertically the individual demand curves for that good. The resulting total demand curve indicates the collective willingness to pay for (or marginal benefit of) any given amount of the public good.

The optimal quantity of a public good occurs where the society's willingness to pay for the last unit—the marginal benefit of the good—equals the marginal cost of the good.

LO4.4 Explain how positive and negative externalities cause under- and overallocations of resources.

Externalities, or spillovers, are costs or benefits that accrue to someone other than the immediate buyer or seller. Such costs or benefits are not captured in market demand or supply curves and therefore cause the output of certain goods to vary from society's optimal output. Negative externalities (or spillover costs or external costs) result in an overallocation of resources to a particular product. Positive externalities (or spillover benefits or external benefits) are accompanied by an underallocaton of resources to a particular product.

Direct controls and specific taxes can improve resource allocation in situations where negative externalities affect many people and community resources. Both direct controls (for example, smokestack emission standards) and specific taxes (for example, taxes on firms producing toxic chemicals) increase production costs and hence product price. As product price rises, the externality, overallocation of resources, and efficiency loss are reduced since less of the output is produced.

Government can correct the underallocation of resources and therefore the efficiency losses that result from positive externalities in a particular market either by subsidizing consumers (which increases market demand) or by subsidizing producers (which increases market supply). Such subsidies increase the equilibrium output, reducing or eliminating the positive externality and consequent underallocation of resources and efficiency loss.

The Coase theorem suggests that under the right circumstances private bargaining can solve externality problems. Thus, government intervention is not always needed to deal with externality problems.

LO4.5 Show why we normally won't want to pay what it would cost to eliminate every last bit of a negative externality such as air pollution.

The socially optimal amount of externality abatement occurs where society's marginal cost and marginal benefit of reducing the externality are equal. With pollution, for example, this optimal amount of pollution abatement is likely to be less than a 100 percent reduction. Changes in technology or changes in society's attitudes toward pollution can affect the optimal amount of pollution abatement.

Market failures present government with opportunities to improve the allocation of society's resources and thereby enhance society's total well-being. But even when government correctly identifies the existence and cause of a market failure, political pressures may make it difficult or impossible for government officials to implement a proper solution.

TERMS AND CONCEPTS

market failures
demand-side market failures
supply-side market failures
consumer surplus
producer surplus
efficiency losses (or deadweight losses)
private goods
rivalry
excludability
public goods
nonrivalry
nonexcludability

free-rider problem
cost-benefit analysis
marginal-cost–marginal-benefit rule
quasi-public goods
externality
Coase theorem
optimal reduction of an externality

DISCUSSION QUESTIONS

1. Explain the two causes of market failures. Given their definitions, could a market be affected by both types of market failures simultaneously? **LO4.1**
2. Use the ideas of consumer surplus and producer surplus to explain why economists say competitive markets are efficient. Why are below- or above-equilibrium levels of output inefficient, according to these two sets of ideas? **LO4.2**
3. What are the two characteristics of public goods? Explain the significance of each for public provision as opposed to private provision. What is the free-rider problem as it relates to public goods? Is U.S. border patrol a public good or a private good? Why? How about satellite TV? Explain. **LO4.3**
4. What divergences arise between equilibrium output and efficient output when (a) negative externalities and (b) positive externalities are present? How might government correct these divergences? Cite an example (other than the text examples) of an external cost and an external benefit. **LO4.4**
5. Why are spillover costs and spillover benefits also called negative and positive externalities? Show graphically how a tax can correct for a negative externality and how a subsidy to producers can correct for a positive externality. How does a subsidy to consumers differ from a subsidy to producers in correcting for a positive externality? **LO4.4**
6. An apple grower's orchard provides nectar to a neighbor's bees, while the beekeeper's bees help the apple grower by pollinating his apple blossoms. Use Figure 4.6b to explain why this situation of dual positive externalities might lead to an underallocation of resources to both apple growing and beekeeping. How might this underallocation get resolved via the means suggested by the Coase theorem? **LO4.4**
7. The LoJack car recovery system allows the police to track stolen cars. As a result, they not only recover 90 percent of LoJack-equipped cars that are stolen but also arrest many auto thieves and shut down many "chop shops" that take apart stolen vehicles to get at their used parts. Thus, LoJack provides both private benefits and positive externalities. Should the government consider subsidizing LoJack purchases? **LO4.4**
8. Explain why zoning laws, which allow certain land uses only in specific locations, might be justified in dealing with a problem of negative externalities. Explain why in areas where buildings sit close together tax breaks to property owners for installing extra fire prevention equipment might be justified in view of positive externalities. Explain why excise taxes on beer might be justified in dealing with a problem of external costs. **LO4.5**
9. **LAST WORD** Distinguish between a carbon-tax and a cap-and-trade strategy for reducing carbon dioxide and other so-called greenhouse gases (that are believed by many scientists to be causing global warming). Which of the two strategies do you think would have the most political support in an election in your home state? Explain your thinking.

REVIEW QUESTIONS

1. Draw a supply and demand graph and identify the areas of consumer surplus and producer surplus. Given the demand curve, what impact will an increase in supply have on the amount of consumer surplus shown in your diagram? Explain why. **LO4.2**
2. Assume that candle wax is traded in a perfectly competitive market in which the demand curve captures buyers' full willingness to pay while the supply curve reflects all production costs. For each of the following situations, indicate whether the total output should be increased, decreased, or kept the same in order to achieve allocative and productive efficiency. **LO4.2**
   a. Maximum willingness to pay exceeds minimum acceptable price.
   b. MC > MB.
   c. Total surplus is at a maximum.
   d. The current quantity produced exceeds the market equilibrium quantity.
3. Efficiency losses ____________________. **LO4.2**
   a. Are not possible if suppliers are willing to produce and sell a product.
   b. Can only result from underproduction.
   c. Can only result from overproduction.
   d. None of the above.
4. Draw a production possibilities curve with public goods on the vertical axis and private goods on the horizontal axis. Assuming the economy is initially operating on the curve, indicate how the production of public goods might be increased. How might the output of public goods be increased

if the economy is initially operating at a point inside the curve? **LO4.3**

5. Use the distinction between the characteristics of private and public goods to determine whether the following should be produced through the market system or provided by government: (a) French fries, (b) airport screening, (c) court systems, (d) mail delivery, and (e) medical care. State why you answered as you did in each case. **LO4.3**
6. Match each of the following characteristics or scenarios with either the term *negative externality* or the term *positive externality*. **LO4.4**
   - **a.** Overallocation of resources.
   - **b.** Tammy installs a very nice front garden, raising the property values of all the other houses on her block.
   - **c.** Market demand curves are too far to the left (too low).
   - **d.** Underallocation of resources.
   - **e.** Water pollution from a factory forces neighbors to buy water purifiers.
7. Use marginal cost/marginal benefit analysis to determine if the following statement is true or false: "The optimal amount of pollution abatement for some substances, say, dirty water from storm drains, is very low; the optimal amount of abatement for other substances, say, cyanide poison, is close to 100 percent." **LO4.5**

PROBLEMS

1. Refer to Table 4.1. If the six people listed in the table are the only consumers in the market and the equilibrium price is \$11 (not the \$8 shown), how much consumer surplus will the market generate? **LO4.2**
2. Refer to Table 4.2. If the six people listed in the table are the only producers in the market and the equilibrium price is \$6 (not the \$8 shown), how much producer surplus will the market generate? **LO4.2**
3. Look at Tables 4.1 and 4.2 together. What is the total surplus if Bob buys a unit from Carlos? If Barb buys a unit from Courtney? If Bob buys a unit from Chad? If you match up pairs of buyers and sellers so as to maximize the total surplus of all transactions, what is the largest total surplus that can be achieved? **LO4.2**
4. **ADVANCED ANALYSIS** Assume the following values for Figures 4.4a and 4.4b. Q_1 = 20 bags. Q_2 = 15 bags. Q_3 = 27 bags. The market equilibrium price is \$45 per bag. The price at *a* is \$85 per bag. The price at *c* is \$5 per bag. The price at *f* is \$59 per bag. The price at *g* is \$31 per bag. Apply the formula for the area of a triangle (Area = ½ × Base × Height) to answer the following questions. **LO4.2**
   - **a.** What is the dollar value of the total surplus (producer surplus plus consumer surplus) when the allocatively efficient output level is being produced? How large is the dollar value of the consumer surplus at that output level?
   - **b.** What is the dollar value of the deadweight loss when output level Q_2 is being produced? What is the total surplus when output level Q_2 is being produced?
   - **c.** What is the dollar value of the deadweight loss when output level Q_3 is produced? What is the dollar value of the total surplus when output level Q_3 is produced?
5. On the basis of the three individual demand schedules in the following table, and assuming these three people are the only ones in the society, determine (a) the market demand schedule on the assumption that the good is a private good and (b) the collective demand schedule on the assumption that the good is a public good. **LO4.3**

P	Q_d (D_1)	Q_d(D_2)	Q_d(D_3)
\$8	0	1	0
7	0	2	0
6	0	3	1
5	1	4	2
4	2	5	3
3	3	6	4
2	4	7	5
1	5	8	6

6. Use your demand schedule for a public good, determined in problem 5, and the following supply schedule to ascertain the optimal quantity of this public good. **LO4.3**

P	Q_s
\$19	10
16	8
13	6
10	4
7	2
4	1

7. Look at Tables 4.1 and 4.2, which show, respectively, the willingness to pay and willingness to accept of buyers and sellers of bags of oranges. For the following questions, assume that the equilibrium price and quantity will depend on the indicated changes in supply and demand. Assume that the only market participants are those listed by name in the two tables. **LO4.4**
   - **a.** What are the equilibrium price and quantity for the data displayed in the two tables?

b. What if, instead of bags of oranges, the data in the two tables dealt with a public good like fireworks displays? If all the buyers free ride, what will be the quantity supplied by private sellers?

c. Assume that we are back to talking about bags of oranges (a private good), but that the government has decided that tossed orange peels impose a negative externality on the public that must be rectified by imposing a $2-per-bag tax on sellers. What is the new equilibrium price and quantity? If the new equilibrium quantity is the optimal quantity, by how many bags were oranges being overproduced before?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

Information Failures

LO4.6 Describe how information failures may justify government intervention in some markets.

This chapter discussed the two most common types of market failure, public goods and externalities. But there is also another, subtler, type of market failure. This one results when either buyers or sellers have incomplete or inaccurate information and their cost of obtaining better information is prohibitive. Technically stated, this market failure occurs because of **asymmetric information**—unequal knowledge possessed by the parties to a market transaction. Buyers and sellers do not have identical information about price, quality, or some other aspect of the good or service.

Sufficient market information is normally available to ensure that goods and services are produced and purchased efficiently. But in some cases inadequate information makes it difficult to distinguish trustworthy from untrustworthy sellers or trustworthy from untrustworthy buyers. In these markets, society's scarce resources may not be used efficiently, thus implying that the government should intervene by increasing the information available to the market participants. Under rare circumstances the government may itself supply a good for which information problems have prohibited efficient production.

ORIGIN OF THE IDEA

O4.4
Information failures

Inadequate Buyer Information about Sellers

Inadequate information among buyers about sellers and their products can cause market failure in the form of underallocation of resources. Two examples will help you understand this point.

Example: Gasoline Market

Assume an absurd situation: Suppose there is no system of weights and measures established by law, no government inspection of gasoline pumps, and no law against false advertising. Each gas station can use whatever measure it chooses; it can define a gallon of gas as it pleases. A station can advertise that its gas is 87 octane when in fact it is only 75. It can rig its pumps to indicate that it is providing more gas than the amount being delivered.

Obviously, the consumer's cost of obtaining reliable information under such chaotic conditions is exceptionally high, if not prohibitive. Customers or their representatives would have to buy samples of gas from various gas stations, have them tested for octane level, and test the accuracy of calibrations at the pump. And these activities would have to be repeated regularly, since a station owner could alter the product quality and the accuracy of the pump at will.

Because of the high cost of obtaining information about the seller, many consumers would opt out of this chaotic market. One tankful of a 50 percent mixture of gasoline and water would be enough to discourage most motorists from further driving. More realistically, the conditions in this market would encourage consumers to vote for political candidates who promise to provide a government solution. The oil companies and honest gasoline stations would most likely welcome government intervention. They would realize that accurate information, by enabling this market to work, would expand their total sales and profits.

The government has in fact intervened in the market for gasoline and other markets with similar potential information difficulties. It has established a system of weights and measures, employed inspectors to check the accuracy of gasoline pumps, and passed laws against fraudulent claims and misleading advertising. Clearly, these government activities have produced net benefits for society.

Example: Licensing of Surgeons

Suppose now that anyone could hang out a shingle and claim to be a surgeon, much as anyone can become a house painter. The market would eventually sort out the true surgeons from those who are "learning by doing" or are fly-by-night operators who move into and out of an area. As people died from unsuccessful surgeries, lawsuits for malpractice eventually would identify and eliminate most of the medical impostors. People needing surgery for themselves or their loved ones could obtain information from newspaper reports, Internet sites, or people who have undergone similar operations.

But this process of obtaining information for those needing surgery would take considerable time and would impose unacceptably high human and economic costs.

There is a fundamental difference between getting an amateurish paint job on one's house and being on the receiving end of heart surgery by a bogus physician. The marginal cost of obtaining information about sellers in the surgery market would be excessively high. The risk of proceeding without good information would result in much less surgery than desirable—an underallocation of resources to surgery.

The government has remedied this market failure through a system of qualifying tests and licensing. The licensing provides consumers with inexpensive information about a service they only infrequently buy. The government has taken a similar role in several other areas of the economy. For example, it approves new medicines, regulates the securities industry, and requires warnings on containers of potentially hazardous substances. It also requires warning labels on cigarette packages and disseminates information about communicable diseases. And it issues warnings about unsafe toys and inspects restaurants for health-related violations.

Inadequate Seller Information about Buyers

Just as inadequate information about sellers can keep markets from achieving economic efficiency, so can inadequate information about buyers. The buyers may be consumers who buy products or firms that buy resources.

Moral Hazard Problem

Private markets may underallocate resources to a particular good or service for which there is a severe **moral hazard problem.** The moral hazard problem is the tendency of one party to a contract or agreement to alter her or his behavior, after the contract is signed, in ways that could be costly to the other party.

Suppose a firm offers an insurance policy that pays a set amount of money per month to people who suffer divorces. The attractiveness of such insurance is that it would pool the economic risk of divorce among thousands of people and, in particular, would protect spouses and children from the economic hardship that divorce often brings. Unfortunately, the moral hazard problem reduces the likelihood that insurance companies can profitably provide this type of insurance.

After taking out such insurance, some people would alter their behavior in ways that impose heavy costs on the insurer. For example, married couples would have less of an incentive to get along and to iron out marital difficulties. At the extreme, some people might be motivated to obtain a divorce, collect the insurance, and then continue to live together. Such insurance could even promote divorce, the very outcome that it is intended to protect against. The moral hazard problem would force the insurer to charge such high premiums for this insurance that few policies would be bought. If the insurer could identify in advance those people most prone to alter their behavior, the firm could exclude them from buying it. But the firm's marginal cost of getting such information is too high compared with the marginal benefit. Thus, this market would fail.

Although divorce insurance is not available in the marketplace, society recognizes the benefits of insuring against the hardships of divorce. It has corrected for this underallocation of "hardship insurance" through child-support laws that dictate payments to the spouse who retains the children, when the economic circumstances warrant them. Alimony laws also play a role.

The government also supplies "divorce insurance" of a sort through the Temporary Assistance for Needy Families (TANF) program. Though aimed at helping poor children in general rather than children of divorce specifically, parents with children can receive TANF payments if they are left destitute by divorce. Because government does not have to earn a profit when supplying services, it can offer this type of "divorce insurance" despite the fact that it, too, may be susceptible to the moral hazard problem.

The moral hazard problem is also illustrated in the following statements:

- Drivers may be less cautious because they have car insurance.
- Medical malpractice insurance may increase the amount of malpractice.
- Guaranteed contracts for professional athletes may reduce the quality of their performance.
- Unemployment compensation insurance may lead some workers to shirk.
- Government insurance on bank deposits may encourage banks to make risky loans.

Adverse Selection Problem

Another information problem resulting from inadequate information about buyers is the **adverse selection problem.** This problem arises when information known by the first party to a contract or agreement is not known by the second and, as a result, the second party incurs major costs. Unlike the moral hazard problem, which arises after a person signs a contract, the adverse selection problem arises at the time a person signs a contract.

In insurance, the adverse selection problem is that people who are most likely to need insurance payouts are those who buy insurance. For example, those in poorest health will seek to buy the most generous health insurance policies. Or, at the extreme, a person planning to hire an arsonist to "torch" his failing business has an incentive to buy fire insurance.

Our hypothetical divorce insurance sheds further light on the adverse selection problem. If the insurance firm sets the premiums on the basis of the average divorce rate, many married couples who are about to obtain a divorce will buy insurance. An insurance premium based on average probabilities will make a great buy for those about to get divorced. Meanwhile, those in highly stable marriages will not buy it.

The adverse selection problem thus tends to eliminate the pooling of low and high risks, which is the basis of profitable insurance. Insurance rates then must be so high that few people would want to (or be able to) buy such insurance.

Where private firms underprovide insurance because of information problems, the government often establishes some type of social insurance. It can require that everyone in a particular group take the insurance and thereby can overcome the adverse selection problem. Example: Although the Social Security system in the United States is partly insurance and partly an income transfer program, in its broadest sense it is insurance against poverty during old age. The Social Security program requires nearly universal participation: People who are most likely to need the minimum benefits that Social Security provides are automatically participants in the program. So, too, are those not likely to need the benefits. Consequently, no adverse selection problem emerges.

Qualification

Households and businesses have found many ingenious ways to overcome information difficulties without government intervention. For example, many firms offer product warranties to overcome the lack of information about themselves and their products. Franchising also helps overcome this problem. When you visit a Wendy's or a Marriott, you know what you are going to get, as opposed to stopping at Slim's Hamburger Shop or the Triple Six Motel.

Also, some private firms and organizations specialize in providing information to buyers and sellers. *Consumer Reports*, *Mobil Travel Guide*, and numerous Internet sites provide product information; labor unions collect and disseminate information about job safety; and credit bureaus provide information about credit histories and past bankruptcies to lending institutions and insurance companies. Brokers, bonding agencies, and intermediaries also provide information to clients.

Economists agree, however, that the private sector cannot remedy all information problems. In some situations, government intervention is desirable to promote an efficient allocation of society's scarce resources.

APPENDIX SUMMARY

LO4.6 Describe how information failures may justify government intervention in some markets.

Asymmetric information occurs when buyers and sellers do not have the same information about a product. It is a source of potential market failure, causing society's scarce resources to be allocated inefficiently.

Asymmetric information can cause a market to fail if the party that has less information decides to withdraw from the market because it fears that its lack of knowledge may be exploited by the party that has more information.

If the party that has less information reduces its participation in a market, the reduction in the size of the market may cause an underallocation of resources to the product produced for the market.

The moral hazard problem is the tendency of one party to a contract or agreement to alter its behavior in ways that are costly to the other party; for example, a person who buys insurance may willingly incur added risk.

The adverse selection problem arises when one party to a contract or agreement has less information than the other party and incurs a cost because of that asymmetrical information. For example, an insurance company offering "no medical-exam-required" life insurance policies may attract customers who have life-threatening diseases.

APPENDIX TERMS AND CONCEPTS

asymmetric information

moral hazard problem

adverse selection problem

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Because medical records are private, an individual applying for health insurance will know more about his own health conditions than will the insurance companies to which he is applying for coverage. Is this likely to increase or decrease the insurance premium that he will be offered? Why? **LO4.6**
2. Why is it in the interest of new homebuyers and builders of new homes to have government building codes and building inspectors? **LO4.6**
3. Place an "M" beside the items in the following list that describe a moral hazard problem and an "A" beside those that describe an adverse selection problem. **LO4.6**
   - **a.** A person with a terminal illness buys several life insurance policies through the mail.
   - **b.** A person drives carelessly because she has automobile insurance.
   - **c.** A person who intends to torch his warehouse takes out a large fire insurance policy.
   - **d.** A professional athlete who has a guaranteed contract fails to stay in shape during the off season.
   - **e.** A woman who anticipates having a large family takes a job with a firm that offers exceptional childcare benefits.

APPENDIX REVIEW QUESTIONS

1. People drive faster when they have auto insurance. This is an example of: **LO4.6**
   - **a.** Adverse selection.
   - **b.** Asymmetric information.
   - **c.** Moral hazard.
2. Government inspectors who check on the quality of services provided by retailers as well as government requirements for licensing in various professions are both attempts to resolve: **LO4.6**
   - **a.** The moral hazard problem.
   - **b.** The asymmetric information problem.
3. True or False: A market may collapse and have relatively few transactions between buyers and sellers if buyers have more information than sellers. **LO4.6**

APPENDIX PROBLEMS

1. Consider a used car market with asymmetric information. The owners of used cars know what their vehicles are worth but have no way of credibly demonstrating those values to potential buyers. Thus, potential buyers must always worry that the used car they are being offered may be a low quality "lemon." **LO4.6**
   - **a.** Suppose that there are equal numbers of good and bad used cars in the market and that good used cars are worth \$13,000 while bad used cars are worth \$5,000. What is the average value of a used car?
   - **b.** By how much does the average value exceed the value of a bad used car? By how much does the value of a good used car exceed the average value?
   - **c.** Would a potential seller of a good used car be willing to accept the average value as payment for her vehicle?
   - **d.** If a buyer negotiates with a seller to purchase the seller's used car for a price equal to the average value, is the car more likely to be good or bad?
   - **e.** Will the used-car market come to feature mostly—if not exclusively—lemons? How much will used cars end up costing if all the good cars are withdrawn?

CHAPTER **5**

Government's Role and Government Failure

Learning Objectives

LO5.1 **Describe how government's power to coerce can be economically beneficial and list some of the difficulties associated with managing and directing the government.**

LO5.2 **Discuss "government failure" and explain why it happens.**

LO5.3 **(Appendix) Explain the difficulties of conveying economic preferences through majority voting.**

Governments in market economies perform several economic tasks. As discussed in various places in the book, these include promoting production and trade by defining property rights, enforcing contracts, and settling disputes; enforcing laws designed to maintain competition; redistributing income via taxes and transfers; reallocating resources by producing public goods and intervening to correct negative and positive externalities; and promoting economic growth and full employment.

In this chapter, we deepen our understanding of government's role in the market economy by examining some of the difficulties that democratic governments face when making specific laws related to the economy. We will find that governments sometimes pursue policies for which costs outweigh benefits. These inefficient outcomes happen often enough that we need to be just as vigilant in looking for instances of *government failure* as we are in looking for instances of *market failure*.

Government's Economic Role

LO5.1 Describe how government's power to coerce can be economically beneficial and list some of the difficulties associated with managing and directing the government.

As discussed in Chapter 2, the U.S. economy is a *market system* that uses mostly markets and prices to coordinate and direct economic activity. But the government also has a prominent role in how the economy functions. Among other things, the government sets the laws governing economic activity, provides goods and services that would otherwise be underproduced by private firms, and modifies the distribution of income. The government also promotes both economic stability and economic growth.

Government's Right to Coerce

One key difference between the economic activities of government and those of private firms and individuals is that government possesses the legal right to force people to do things. Whereas private-sector economic activities consist primarily of voluntary transactions, government has the legal right to enforce involuntary transactions. Among other things, the government can put you in jail if you do not pay your taxes, fine you if you violate pollution laws, jail you if you commit fraud, and remove your business license if you violate health and safety regulations.

Force and Economic Efficiency From an economic perspective, the government's ability to force people to do things can be quite beneficial because it can be used to increase economic efficiency.

Correcting for Market Failures Consider public goods and externalities. As discussed in Chapter 4, these market failures cause resource misallocations. When it comes to both public goods and products offering positive externalities, private producers fail to produce enough output because it is impossible to charge many of the beneficiaries for the benefits that they receive from the producers' products. In such cases, the government can improve economic efficiency by using involuntarily collected tax money to subsidize production.

By contrast, products that generate negative externalities are overproduced by the private sector because many of their costs are borne by third parties rather than by their producers. The government can reduce that overproduction and improve economic efficiency by using involuntary policies such as direct controls, pollution taxes, and cap-and-trade schemes to force producers to bear higher costs.

Reducing Private-Sector Economic Risks Government's ability to force people to do things is also crucial in reducing private-sector economic risks. To begin with, the government helps to ensure that only mutually agreeable transactions take place by making blackmail, extortion, and other forms of private coercion illegal. The government also uses its legal powers to outlaw various forms of theft, deception, and discrimination as well as restraints on trade, price-fixing, and refusal to honor a contract.

These limitations encourage economic activity by giving greater security to both individuals and firms. Because they know that the government will use its massive resources to arrest and punish those who break the law, they know that other individuals and firms are less likely to try to take advantage of them. That reduction in risk encourages higher levels of investment, the formation of more new businesses, and the introduction of more new goods and services. In economic terminology, both allocative and productive efficiency increase.

The Problem of Directing and Managing Government

As just discussed, the government can substantially improve allocative and productive efficiency if it directs its awesome coercive powers toward rectifying market failures and providing a low-risk economic environment for the private sector. However, it has only been in recent centuries that democratic political institutions have been able to tame government and direct it toward those goals. Until that happened, most governments were tyrannical, with their powers almost always directed toward enriching the small minority that controlled each government.

Because modern democratic governments serve much broader constituencies, they are much more likely to pursue economic policies with widespread social benefits. Their ability to deliver economically optimal outcomes is hindered, however, by the wide variety of government failures that this chapter will discuss in detail.

But before discussing them, it will be useful to first point out that governing a nation is not easy. In particular, governments face the daunting challenge of organizing millions of employees to carry out thousands of tasks—everything from cleaning sewers to researching cures for cancer to delivering the mail. An understanding

CONSIDER THIS . . .

Does Big Government Equal Bad Government?

You will sometimes hear politicians (and maybe your grumpy uncle) complaining about Big Government. Their implication is that large government initiatives are inherently inefficient or incompetent.

Since economics is focused on efficiency, you might wonder where economists stand on the subject.

The answer is that economists focus not on bigness or smallness *per se,* but on marginal benefit (MB) and marginal cost (MC). Spending should be increased up to the point where MB = MC. For some programs, that will be a small dollar amount. For other programs, that will be a large dollar amount.

Thus, economists don't see much point in having an abstract debate over "big government" versus "small government." What matters is allocative and productive efficiency and directing government's limited resources toward the programs that generate the largest net benefits for society.

From that vantage point, we should not condemn large government programs just for being large. We must first compare MB with MC. Only if MB < MC should large programs be reduced or eliminated.

of those challenges and complexities will give you a better sense of how well most governments manage to do *despite* all of the problems associated with government failure.

No Invisible Hand Government economic polices are not self-correcting. Unlike the private sector—where competitive forces and Adam Smith's "invisible hand" help to automatically direct resources to their best uses—poorly designed government policies can misallocate resources indefinitely unless active steps are taken by legislators or administrators.

Massive Size and Scope Identifying and correcting inefficient government policies is hampered by government's massive size and scope. Consider the U.S. federal government. In 2010, it had 4.4 million employees spread over 500 agencies that were collectively charged with enforcing hundreds of thousands of pages of laws and regulations while attempting to wisely spend $3.4 trillion.

The Need for Bureaucracy By law, those 4.4 million federal employees are ultimately supervised and directed by just 536 elected officials: one president, 435 representatives, and 100 senators. Since 536 elected officials could never hope to directly supervise 4.4 million people, governments rely on many layers of supervisors and supervisors-of-supervisors to manage the government's affairs. They collectively form a massive, hierarchical, many-layered bureaucracy.

The Need for Paperwork and Inflexibility To make sure that laws are uniformly enforced and do not vary at the whim of individual bureaucrats, the bureaucracy is regulated by detailed rules and regulations governing nearly every possible action that any individual bureaucrat might be called upon to make. These rules and regulations ensure that laws and regulations are uniformly applied. But they do so at the cost of massive amounts of paperwork and an inability to expeditiously process nonroutine situations and requests.

The Information Aggregation Problem Because of their massive size and scope, bureaucracies have difficulty with effectively aggregating and conveying information from their bottom layers to their top layers. As a result, top officials will tend to make many inefficient choices because they do not have enough information to sensibly compare the marginal benefits and marginal costs of individual programs and because they are unable to comprehensively assess opportunity costs and where to best spend funds across the wide variety of programs run by the government.

Lack of Accountability Governments also struggle with accountability. Democratic elections do take place for the elected officials at the top, but because the government undertakes so many activities simultaneously, it is difficult for the electorate to know the details of even a small fraction of what the government is up to at any particular time. As a result, hundreds or even thousands of individual programs may be poorly run without affecting the reelection chances of the incumbent politicians who are supposed to be supervising everything.

Within the bureaucracy itself, individual accountability is also hard to enforce because most bureaucrats have civil service protections that effectively guarantee them a job for life. Those protections reduce corruption by shielding bureaucrats from political pressures. But they also severely constrain the ability of elected officials to hold individual bureaucrats personally responsible for bad decisions.

QUICK REVIEW 5.1

- Government's ability to enforce nonvoluntary transactions can improve economic outcomes by compensating for resource misallocations and by providing a low-risk economic environment for individuals and firms.
- Government economic actions are not automatically self-correcting (as with the "invisible hand" in competitive markets.)
- Democratic governments face several challenges in directing and supervising government's actions, including inflexibility, information aggregation, comparing marginal costs with marginal benefits, assessment of opportunity costs, and accountability.

Government Failure

LO5.2 Discuss "government failure" and explain why it happens.

The term **government failure** refers to economically inefficient outcomes caused by shortcomings in the public sector. One cause of government failure is the voting problems that we discuss at length in this chapter's appendix. But government failures caused by voting problems are somewhat unique in that they are driven by a lack of information about voter preferences. By contrast, most instances of government failure happen *despite* government officials knowing what voters prefer.

In these situations, government failures occur because the incentive structures facing government officials lead them to either put their own interests ahead of voter interests or to put the interests of a minority of voters ahead of those of the majority of voters. Let's examine what economic theory has to say about these situations.

Representative Democracy and the Principal-Agent Problem

Our system of representative democracy has the advantage of allowing us to elect full-time representatives who can specialize in understanding the pros and cons of different potential laws and who have more time to digest their details than the average citizen. But the system also suffers from principal-agent problems.

Principal-agent problems are conflicts that arise when tasks are delegated by one group of people (principals) to another group of people (agents). The conflicts arise because the interests of the agents may not be the same as the interests of the principals, so that the agents may end up taking actions that are opposed by the principals whom they are supposed to be representing.

In the business world, principal-agent problems often arise when the company's managers (the agents) take actions that are not in the best interests of the company's shareholders (the principals). Examples include the managers spending huge amounts of company money on executive jets and lavish offices or holding meetings at expensive resorts. These luxuries are obviously enjoyable to managers but are, of course, not in the best interest of shareholders because the money spent on them could either be reinvested back into the firm to increase future profits or paid out to shareholders immediately as dividends. But to the extent that managers are free to follow their own interests rather than those of their shareholders, they may indeed take these and other actions that are not in the better interests of their shareholders. Hence the conflicts.

In a representative democracy, principal-agent problems often arise because politicians have goals such as reelection that may be inconsistent with pursuing the best interests of their constituents. Indeed, casual reflection suggests that "sound economics" and "good politics" often differ. Sound economics calls for the public sector to pursue various programs as long as marginal benefits exceed marginal costs. Good politics, however, suggests that politicians support programs and policies that will maximize their chances of getting reelected. The result may be that the government will promote the goals of groups of voters that have special interests to the detriment of the larger public. Economic inefficiency is the likely outcome.

Special-Interest Effect Efficient public decision making is often impaired by the **special-interest effect.** This is any outcome of the political process whereby a small number of people obtain a government program or policy that gives them large gains at the expense of a much greater number of persons who individually suffer small losses.

The small group of potential beneficiaries is well informed and highly vocal on the issue in question, and they press politicians for approval. The large number of people facing the very small individual losses, however, are generally uninformed on the issue. Politicians feel they will lose the campaign contributions and votes of the small special-interest group that backs the issue if they legislate against it but will lose very little support from the large group of

CONSIDER THIS . . .

Mohair and the Collective-Action Problem

Smaller groups can sometimes achieve political victories against larger groups by taking advantage of the **collective-action problem**—the fact that larger groups are more difficult to organize and motivate than smaller groups.

Larger groups are harder to organize and motivate for two main reasons. First, the larger the group, the smaller each member's share of the benefits if the group gets its way. Second, the larger the group, the higher its organizing costs, as it will have to contact and recruit large numbers of strangers via e-mails, telephone calls, and mass mailings.

Smaller groups can take advantage of these difficulties and generally get their way against larger groups as long as they are pressing for policies that only cause small amounts of harm to the members of the larger groups.

Consider the infamous subsidy for mohair, the wool produced by Angora goats. Each year the federal government provides millions of dollars in subsidized loans to Angora goat farmers in Texas, Arizona, and New Mexico. The federal government began the subsidy in the late 1940s to ensure a large supply of insulation for the jackets needed to keep pilots and other crew members warm in the unheated airplanes used during that period.

The mohair subsidy should have ended in the 1950s when heated cabins were developed, but it survives because it costs taxpayers only a few cents each. This means that it would cost them more to organize and defeat the mohair subsidy than they would save by having the subsidy terminated.

More generally, the collective-action problem explains why nearly every example of the special-interest effect is characterized by "concentrated benefits and diffuse costs." Concentrated benefits make proponents easy to organize, while diffuse costs make opponents difficult to organize.

uninformed voters, who are likely to evaluate the politicians on other issues of greater importance to them.

The special-interest effect is also evident in so-called *pork-barrel politics*, a means of securing a government project that yields benefits mainly to a single political district and its political representative. In this case, the special-interest group comprises local constituents, while the larger group consists of relatively uninformed taxpayers scattered across a much larger geographic area. Politicians clearly have a strong incentive to secure government projects ("pork") for their local constituents. Such projects win political favor because they are highly valued by constituents and the costs are borne mainly by taxpayers located elsewhere.

At the federal level, pork-barrel politics often consist of congressional members inserting specific provisions that authorize spending for local projects (that will benefit only local constituents) into comprehensive legislation (that is supposed to be about making laws for the entire country). Such narrow, specifically designated authorizations of expenditure are called **earmarks.** In 2012, legislation contained 152 such earmarks, totaling $3.3 billion. These earmarks enable senators and representatives to provide benefits to in-state firms and organizations without subjecting the proposals to the usual evaluation and competitive bidding. Although some of the earmarked projects deliver benefits that exceed costs, many others are questionable, at best. These latter expenditures very likely reallocate some of society's scarce resources from higher-valued uses to lower-valued uses. Moreover, logrolling, discussed in the chapter appendix, typically enters the picture. "Vote for my special local project and I will vote for yours" becomes part of the overall strategy for securing "pork" and remaining elected.

Finally, a politician's inclination to support the smaller group of special beneficiaries is enhanced because special-interest groups are often quite willing to help finance the campaigns of "right-minded" politicians and politicians who "bring home the pork." The result is that politicians may support special-interest programs and projects that cannot be justified on economic grounds.

Rent-Seeking Behavior The appeal to government for special benefits at taxpayers' or someone else's expense is called **rent seeking.** The term "rent" in "rent seeking" is used loosely to refer to any payment in excess of the minimum amount that would be needed to keep a resource employed in its current use. Those engaged in "rent seeking" are attempting to use government influence to get themselves into a situation in which they will get paid more for providing a good or service than the minimum amount you would actually have to pay them to provide that good or service. (These excess, or surplus, payments are akin to *land rent*, which is also a surplus payment.)

Rent seeking goes beyond the usual profit seeking through which firms try to increase their profits by adjusting their output levels, improving their products, and incorporating cost-saving technologies. Rent seeking looks to obtain extra profit or income by influencing government policies. Corporations, trade associations, labor unions, and professional organizations employ vast resources to secure favorable government policies that result in rent—higher profit or income than would otherwise occur. The government is able to dispense such rent directly or indirectly through laws, rules, hiring, and purchases. Elected officials are willing to provide such rent because they want to be responsive to the key constituents who can help them remain in office.

Here are some examples of "rent-providing" legislation or policies: tariffs on foreign products that limit competition and raise prices to consumers; tax breaks that benefit specific corporations; government construction projects that create union jobs but cost more than the benefits they yield; occupational licensing that goes beyond what is needed to protect consumers; and large subsidies to farmers by taxpayers. None of these is justified by economic efficiency.

Clear Benefits, Hidden Costs

Some critics say that vote-seeking politicians will ignore economic rationality by failing to objectively weigh costs and benefits when deciding which programs to support. Because political officeholders must seek voter support every few years, they favor programs that have immediate and clear-cut benefits and vague or deferred costs. Conversely, politicians will reject programs with immediate and easily identifiable costs but with less measurable but very high long-term benefits.

Such biases may lead politicians to reject economically justifiable programs and to accept programs that are economically irrational. Example: A proposal to construct or expand mass-transit systems in large metropolitan areas may be economically rational on the basis of cost-benefit analysis. But if (1) the program is to be financed by immediate increases in highly visible income or sales taxes and (2) benefits will occur only years from now when the project is completed, then the vote-seeking politician may oppose the program.

Assume, on the other hand, that a program of federal aid to municipal police forces is not justifiable on the basis of cost-benefit analysis. But if the cost is paid for from budget surpluses, the program's modest benefits may seem so large that it will gain approval.

Unfunded Liabilities

The political tendency to favor spending priorities that have immediate payouts but deferred costs also leads to many government programs having unfunded liabilities. A government creates an **unfunded liability** when it commits to making a series of future expenditures without simultaneously committing to collect enough tax revenues to pay for those expenditures.

The most famous example of an unfunded liability belongs to the Social Security program, under which the U.S. federal government supplements the incomes of the elderly and the disabled. The government does collect Social Security taxes to help defray the expected future costs of the program, but the current tax rates will not generate nearly enough revenue to pay for all of the expected outlays. In fact, it is estimated that Social Security has an unfunded liability (= total value of spending commitments minus expected value of tax revenues) of $20.5 trillion.

Social Security is not the only major unfunded government liability. Medicare, which provides healthcare to the elderly and disabled in the United States, has an unfunded liability of $4.8 trillion, while state and local governments are estimated to have $4.6 trillion in unfunded retirement and healthcare commitments.

Chronic Budget Deficits

A government runs an annual **budget deficit** whenever its tax revenues are less than its spending during a particular year. To make up for the shortfall, the government must borrow money, usually by issuing bonds. Whatever it borrows in a given year gets added to its overall pile of debt, which is the accumulation of all past budget deficits and budget surpluses.

Many governments run budget deficits year after year. These chronic deficits can be attributed to a pair of conflicting incentives that confront politicians. On the one hand, many government programs are highly popular with voters, so that there is almost always political pressure to either maintain or increase spending. On the other hand, hardly anyone likes paying taxes, so there is almost always political pressure to reduce taxes. Faced with those two conflicting pressures, politicians tend to opt for spending levels that exceed tax revenues.

That may be problematic because chronic deficits can pose several economic challenges, including

- ***Economic Inefficiency*** Deficits may allow the government to control and direct an inefficiently large fraction of the economy's resources. To the extent that deficit spending facilitates an underallocation of resources to the private sector and an overallocation of resources to the government sector, there will be a tendency to underproduce private goods and overproduce public goods. If that occurs, the economy will experience a decrease in both allocative and productive efficiency.
- ***Debt Crises*** A government's accumulated debt level may rise so high that investors lose faith in the government's ability or willingness to repay its debts.

If that happens, the government will find itself in the middle of a **debt crisis,** unable to borrow any more money. Cut off from borrowing, the government will be forced to undertake some combination of drastic spending cuts or massive tax increases. Either of those actions will tend to plunge the economy into a recessionary period in which unemployment rises and output falls.

To prevent politicians from succumbing to voter preferences for deficits, many state and local governments have balanced-budget laws that make deficits illegal. No such law exists at the national level, however. As a result, federal politicians were able to run budget deficits in 47 of the 52 years between 1960 and 2012.

Misdirection of Stabilization Policy

Economies go through alternating periods of expansion and recession. Multiyear periods during which output expands, employment increases, and living standards rise alternate with periods during which output contracts, employment decreases, and living standards fall.

Governments often attempt to smooth out these so-called *business cycles* by using two types of macroeconomic stabilization policy:

- **Fiscal policy** attempts to use changes in tax rates and spending levels to offset the business cycle. For example, if the economy is going into a recessionary period with falling output and rising unemployment, the government may attempt to stimulate the economy by lowering tax rates or increasing government spending. Either action should increase spending on goods and services and consequently induce business to produce more output and hire more workers.
- **Monetary policy** attempts to use changes in interest rates to regulate the economy. In particular, the government can use its control over the money supply to lower interest rates during a recession. The lower interest rates stimulate spending by making it cheaper for individuals and businesses to borrow money to pay for capital goods such as houses, cars, and machinery. As spending on those items increases, firms are induced to produce more output and hire more workers.

Politicization of Fiscal and Monetary Policy Fiscal and monetary policy are both subject to politicization. In the case of fiscal policy, if the economy goes into recession and there are calls to stimulate the economy through lower taxes or increased spending, politicians often spend more time attempting to target any tax cuts or spending increases toward special interests than they do making sure that their fiscal policy actions will actually stimulate the overall economy. The recession also provides political cover for increasing the size of the deficit.

Monetary policy can be similarly politicized, with the biggest problem being that incumbent politicians will want to cut interest rates to boost the economy right before they are up for reelection. That is problematic because monetary stimulus is only helpful if the economy is in recession. If the economy is doing well, monetary stimulus can actually make things worse because it can raise the rate of inflation and drive up prices all over the economy.

To prevent that, most countries have put politically independent central banks in charge of monetary policy. In the United States, the Federal Reserve serves this function. Other top central banks include the Bank of Japan, the Bank of England, and the European Central Bank. Each is run by professional economists who are insulated from political pressures so that they may use their independent expertise and judgment to decide if and when monetary stimulus should be used.

QUICK REVIEW 5.2

- Principal-agent problems are conflicts that occur when the agents who are supposed to be acting in the best interests of their principals instead take actions that help themselves but hurt their principals.
- Because larger groups are more difficult to organize and motivate than smaller groups, special interests can often obtain what they want politically even when what they want is opposed by a majority of voters.
- Rent seeking involves influencing government policies so that one can get paid more for providing a good or service than it costs to produce.
- Political pressures cause politicians to favor policies such as unfunded liabilities and budget deficits that have immediate benefits and delayed costs.

Limited and Bundled Choice

Economic theory points out that the political process forces citizens and their elected representatives to be less selective in choosing public goods and services than they are in choosing private goods and services.

In the marketplace, the citizen as a consumer can exactly satisfy personal preferences by buying certain goods and not buying others. However, in the public sector the citizen as a voter is confronted with, say, only two or three candidates for an office, each representing a different "bundle" of programs (public goods and services). None of these bundles of public goods is likely to fit exactly the preferences of any particular voter. Yet the voter must choose one of them. The candidate who comes closest to voter Smith's preference may endorse national health insurance, increases in Social Security benefits, subsidies to tobacco farmers, and tariffs on imported goods. Smith is likely to vote for that candidate even though Smith strongly opposes tobacco subsidies.

In other words, the voter must take the bad with the good. In the public sector, people are forced to "buy" goods and services they do not want. It is as if, in going to a sporting-goods store, you were forced to buy an unwanted pool cue to get a wanted pair of running shoes. This is a situation where resources are not being used efficiently to satisfy consumer wants. In this sense, the provision of public goods and services is inherently inefficient.

Congress is confronted with a similar limited-choice, bundled-goods problem. Appropriations legislation combines hundreds, even thousands, of spending items into a single bill. Many of these spending items may be completely unrelated to the main purpose of the legislation. Yet congressional representatives must vote on the entire package—yea or nay. Unlike consumers in the marketplace, they cannot be selective.

Bureaucracy and Inefficiency

Some economists contend that public agencies are generally less efficient than private businesses. The reason is not that lazy and incompetent workers somehow end up in the public sector while ambitious and capable people gravitate to the private sector. Rather, it is that the market system creates incentives for internal efficiency that are absent from the public sector. Private enterprises have a clear goal—profit. Whether a private firm is in a competitive or monopolistic market, efficient management means lower costs and higher profit. The higher profit not only benefits the firm's owners but enhances the promotion prospects of the firm's managers. Moreover, part of the managers' pay may be tied to profit via profit-sharing plans, bonuses, and stock options. There is no similar gain to government agencies and their managers—no counterpart to profit—to create a strong incentive to achieve efficiency.

CONSIDER THIS . . .

Unintended Consequences

As explained in Chapters 2 and 4, the "invisible hand" of a properly functioning market will allocate resources to their best uses without anyone being in charge or intentionally aiming for efficiency. By contrast, governments are willful and intentional. They deliberately create and enforce laws to try to make improvements in society. In some cases, however, government actions can have **unintended consequences** that offset some or all of the intended benefits.

- Government fuel-efficiency requirements for automobiles have forced automakers to produce smaller, lighter vehicles. But when smaller, lighter vehicles get into accidents, their occupants are more likely to be killed or severely injured. Some estimates put the death toll at over 120,000 additional deaths in the United States since 1970.
- San Francisco banned plastic grocery bags in 2007. This led to about 5 additional deaths per year from food-borne illnesses because reusable grocery bags almost never get washed out. Drippings from one trip often fester and contaminate whatever they touch on subsequent trips.
- The main point of the 2010 healthcare reform law (commonly known as Obamacare) was to get health insurance coverage for all Americans. To that end, the law required larger companies to either pay for extremely costly health insurance policies for their full-time workers or face massive fines. But since that requirement only applied to full-time workers, many firms responded by cutting a lot of their employees' work hours down from full time to part time. Thus, millions of workers went from lacking health insurance but having a full-time job to still lacking health insurance but only having a part-time job.

The market system imposes a very obvious test of performance on private firms: the test of profit and loss. An efficient firm is profitable and therefore successful; it survives, prospers, and grows. An inefficient firm is unprofitable and unsuccessful; it declines and in time goes out of business. But there is no similar, clear-cut test with which to assess the efficiency or inefficiency of public agencies. How can anyone determine whether a public hydroelectricity provider, a state university, a local fire department, the Department of Agriculture, or the Bureau of Indian Affairs is operating efficiently?

Cynics even argue that a public agency that inefficiently uses its resources is likely to survive and grow! In the private sector, inefficiency and monetary loss lead to the abandonment of certain activities or products or even firms. But the government, they say, does not like to abandon activities in which it has failed. Some suggest that the typical response of the government to a program's failure is to increase its budget and staff. This means that public sector inefficiency just continues on a larger scale.

Furthermore, economists assert that government employees, together with the special-interest groups they serve, often gain sufficient political clout to block attempts to pare down or eliminate their agencies. Politicians who attempt to reduce the size of huge federal bureaucracies such as those relating to agriculture, education, health and welfare, and national defense incur sizable political risk because bureaucrats and special-interest groups will team up to defeat them.

Finally, critics point out that government bureaucrats tend to justify their continued employment by looking for and eventually finding new problems to solve. It is not surprising that social "problems," as defined by government, persist or even expand.

The Last Word at the end of this chapter highlights several recent media-reported examples of the special-interest effect (including earmarks), the problem of limited and bundled choices, and problems of government bureaucracy.

Inefficient Regulation and Intervention

Governments regulate many aspects of the market economy. Examples include health and safety regulations, environmental laws, banking supervision, restrictions on monopoly power, and the imposition of wage and price controls.

These interventions are designed to improve economic outcomes, but several forms of regulation and intervention have been known to generate outcomes that are less beneficial than intended.

Regulatory Capture A government agency that is supposed to supervise a particular industry is said to have suffered from **regulatory capture** if its regulations and enforcement activities come to be heavily influenced by the industry that it is supposed to be regulating.

Regulatory capture is often facilitated by the fact that nearly everyone who knows anything about the details of a regulated industry works in the industry. So when it comes time for the regulatory agency to find qualified people to help write intelligent regulations, it ends up hiring a lot of people from regulated firms. Those individuals bring their old opinions and sympathies with them when they become bureaucrats. As a result, many regulations end up favoring the interests of the regulated firms.

Regulatory Capture in the Railroad Industry The classic example of regulatory capture is that of railroad regulation during the nineteenth and twentieth centuries. In response to public complaints that the nation's railroads were often charging exorbitant rates, the federal government established the Interstate Commerce Commission (ICC) in 1887 as the government agency charged with regulating competition and prices within the railroad industry.

Within a generation, railroad executives had achieved regulatory capture by manipulating the ICC into a policy that simultaneously fixed rates at profitable levels while also eliminating competition between different railroad companies. The public justification for these policies was that competition had to be restricted in order to prevent larger railroads from bankrupting smaller railroads and thereby becoming monopolies that could easily exploit the public. But the railroad industry's true motive was to establish a regulatory regime in which both larger and small railroads were guaranteed steady, competition-free profits.

These days, activists often complain that various government bureaucracies are subject to regulatory capture. At the federal level, complaints are voiced about the Food and Drug Administration's supervision of the pharmaceutical industry, the Securities and Exchange Commission's supervision of Wall Street financial firms, and the Bureau of Land Management's policies with respect to leasing federal lands for oil drilling, mining, and forestry.

Deregulation as an Alternative Economists are divided about the intensity and inefficiency of regulatory capture as well as what to do about it. One potential solution is for the government to engage in **deregulation** and intentionally remove most or even all of the regulations governing an industry.

Deregulation solves the problem of regulatory capture because there is no regulatory agency left to capture. But it only works well in terms of economic efficiency if the deregulated industry becomes competitive and is automatically guided toward allocative and productive efficiency by competitive forces and the invisible hand. If the deregulated industry instead tends toward monopoly or ends up generating substantial negative externalities, continued regulation might be the better option.

Proponents of deregulation often cite the deregulation of interstate trucking, railroads, and airlines in the 1970s and 1980s as examples of competition successfully replacing regulation. They do so because after regulation was removed, robust competition led to lower prices, increased output, and higher levels of productivity and efficiency.

But for government agencies tasked with environmental protection, human safety, and financial regulation, there is less confidence as to whether competitive pressures might be able to replace regulation. For those industries, regulation may always be necessary. If so, then some amount of regulatory capture may always be likely due to the fact that regulated firms will always want to capture their regulators.

Government's Poor Investment Track Record

Governments are often asked to use taxpayer money to directly invest in private businesses that have been unable to secure funding from private sources such as banks. Unfortunately, researchers have found that low and negative rates of return are the norm for government investments. In addition, government funding often allows inefficient firms to persist in operation long after competitive forces would have put them out of operation and freed up their resources for higher-valued projects elsewhere in the economy.

Critics also note that many government investments look like prime examples of rent seeking and the special-interest effect, especially when the firms receiving government investments are found to have made substantial financial contributions to influential politicians. In too many cases, the government's investment decisions appear to be based on political connections rather than on whether specific investments can produce substantial net benefits for society.

Loan Guarantees The government also tends to earn low or negative returns when it subsidizes private-sector investments with **loan guarantees.** The startup company named Solyndra provides a good example of what can go wrong.

The Solyndra Subsidy In 2009, Solyndra was unable to convince private investors to lend it enough money to start producing solar panels with its new technology. The private investors sensibly feared that the company's new technology was too expensive and that its solar panels would not be able to compete with those made by the industry's more established firms.

At that point, Solyndra turned to a federal loan-guarantee program under which the Department of Energy told potential investors that it would cosign any loan taken out by Solyndra and thereby guarantee that if Solyndra went bankrupt, the federal government would use taxpayer money to repay the loan.

With that loan guarantee in place, the otherwise-reluctant private investors were willing to put in $535 million. After all, they had nothing to lose and everything to gain. If Solyndra went bankrupt, they would get their money back from the government. But if Solyndra somehow did well, they would collect substantial returns.

Unfortunately, the investors' original doubts proved to be well founded. Solyndra was unable to compete effectively with incumbent firms and went bankrupt in 2011, leaving taxpayers on the hook for the full $535 million.

Socializing Losses, Privatizing Gains Government loan guarantees can be socially beneficial if they help to increase the production of beneficial products that are being underproduced by the private sector—as would be the case for products that generated positive externalities. But the loan guarantees also provide an inducement toward reckless investing because they remove from private investors any consideration of losses. Indeed, loan guarantees are often criticized for "socializing losses and privatizing gains" because if things go wrong, any losses go to the taxpayer, while if things go well, any profits go to private investors.

In addition, the process by which loan guarantees are awarded is often criticized for being highly politicized and likely to award loan guarantees not to the firms whose projects are the most likely to increase economic efficiency but to those with the best political connections.

On the other hand, there may be legitimate cases where a new technology that would generate net benefits cannot be developed without government loan guarantees, so proponents of loan-guarantee programs argue that the programs should remain in place, but with tight controls against rent seeking and the special-interest effect.

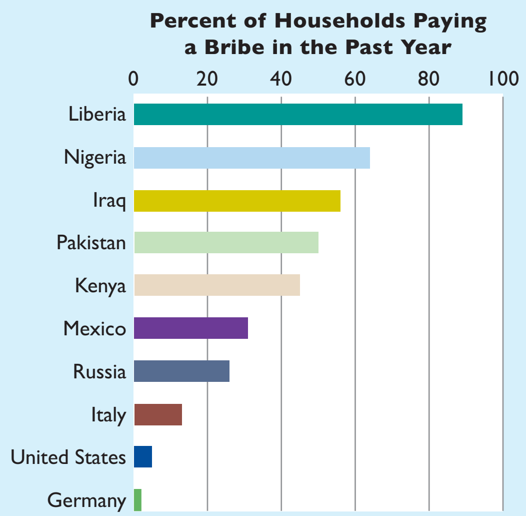
Corruption

Political corruption is the unlawful misdirection of governmental resources or actions that occurs when government officials abuse their entrusted powers for personal gain. For instance, a police supervisor engages in political corruption if she accepts a bribe in exchange for illegally freeing a thief who had been lawfully arrested by another

GLOBAL PERSPECTIVE 5.1

Percentage of Households Paying a Bribe in the Past Year

The Global Corruption Barometer is an international survey that asks individuals about their personal experiences with government corruption. The 2010–2011 survey of 105,507 people in 100 countries included a question that asked participants whether they or anyone in their respective households had paid a bribe in any form during the previous 12 months. Here are the results for 10 selected countries.



Source: Adapted from *Global Corruption Barometer*. Copyright 2011 Transparency International: the global coalition against corruption. Used with permission. For more information, visit **www.transparency.org**.

officer. Similarly, a government bureaucrat engages in political corruption if he refuses to issue a building permit to a homebuilder who is in full compliance with the law unless the homebuilder makes a "voluntary contribution" to the bureaucrat's favorite charity.

While relatively uncommon in the United States, political corruption is a daily reality in many parts of the world, as can be seen in Global Perspective 5.1, which gives the percentages of survey respondents in 15 countries who reported that they or someone else in their respective households paid a bribe during the previous 12 months.

Political corruption comes in two basic forms. In the first, a government official must be bribed to do what he should be doing as part of his job—as with the bureaucrat in our earlier example who demands a bribe to issue a permit to a homebuilder who is in full compliance with the law. In the second, a government official demands a bribe to do something that she is not legally entitled to do—as with the police supervisor in our earlier example who illegally freed a thief.

If a candidate accepts campaign contributions from a special-interest group and then shows subsequent support for that group's legislative goals, has a subtle form of political corruption taken place? While there are strong opinions on both sides of the issue, it is often hard to tell in any particular case whether a special interest's campaign contribution amounts to a bribe. On the one hand, the special interest may indeed be trying to influence the politician's vote. On the other hand, the special interest may simply be trying to support and get elected a person who already sees things their way and who would vote the way they wanted no matter what.

That being said, the impression of impropriety lingers, and so laws have been passed in the United States limiting the amount of money that individuals can donate to specific candidates and making it illegal for certain groups such as companies to donate money directly to individual politicians (as distinct from directing funds toward supporting specific issues or advocacy groups—which is both legal and unrestricted). Proponents of these laws hope that the limitations strike a good balance—allowing contributions to be large enough that individuals and groups can meaningfully support candidates they agree with but keeping contributions small enough that no one individual or group can singlehandedly donate enough money to sway a politician's vote.

Imperfect Institutions

It is possible to argue that the wide variety of criticisms of public sector inefficiency that we have discussed in this chapter are exaggerated and cynical. Perhaps they are. Nevertheless, they do tend to shatter the concept of a benevolent government that responds with precision and efficiency to the wants of its citizens. The market system of the private sector is far from perfectly efficient, and government's economic function is mainly to correct that system's shortcomings. But the public sector is also subject to deficiencies in fulfilling its economic function. "The relevant comparison is not between perfect markets and imperfect governments, nor between faulty markets and all-knowing, rational, benevolent governments, but between inevitably imperfect institutions."[1]

Because markets and governments are both imperfect, it is sometimes difficult to determine whether a particular activity can be performed with greater success in the private

[1]Otto Eckstein, *Public Finance*, 3d ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1973), p. 17.

LAST WORD

"Government Failure" in the News

The Media Continually Report Government Actions That Illustrate Pork-Barrel Politics, Limited and Bundled Choices, or Bureaucratic Inefficiency.

Examples:

- A 2004 spending bill set aside $1 million for the Norwegian American Foundation; $443,000 to develop salmon-fortified baby food; $350,000 for music education programs at the Rock and Roll Hall of Fame in Cleveland; and $250,000 for sidewalks, street furniture, and façade improvements in Boca Raton, Florida. (Associated Press)
- The corporate tax relief bill of 2004 contained 633 pages, with 276 special provisions. Included were provisions that benefited "restaurant owners and Hollywood producers; makers of bows, arrows, tackle boxes, and sonar fish finders; NASCAR track owners; native Alaska whalers; and even importers of Chinese fans." (*The Washington Post*)
- Government investigations determined that millions of dollars of disaster relief for victims of Hurricane Katrina were squandered. For example, investigators discovered that the Federal Emergency Management Agency (FEMA) made payouts on as many as 900,000 claims for disaster relief that contained invalid Social Security numbers or false names and addresses. (*The Seattle Times*)
- The $878 billion American Recovery and Reinvestment Act of 2009 was laden with many dubious spending projects, including $10 million to renovate a train station in Elizabethtown, Pennsylvania, that hadn't been used in 30 years; $1.15 million to build a guardrail for an artificial lake in Woodward, Oklahoma, that had never been filled with water; and an unrequested $587,661 grant that was given to the upscale town of Union, New York, to fight a homeless problem that it didn't have. (*Lancaster Newspapers*, **newson6.com**, *Binghamton Press & Sun Union*)

- The year 2009 also saw Congress approve a $2.5 billion earmark to purchase ten C-17 aircraft despite the Department of Defense adamantly stating that its existing fleet of 205 C-17s was "sufficient to meet the Department's future airlift needs—even under the most stressing situations." (**investinganswers.com**)
- In 2011, Congress funded a sanctuary for white squirrels, an antique bicycle museum, and a giant roadside coffee pot as part of 2011 federal highway spending. It also spent $765,828 to subsidize the construction of an IHOP restaurant and $113,277 to aid in the historical preservation of video games. (*Human Events*, *Washington Examiner*, *Gamasutra*)
- A 2011 audit revealed that the federal government had paid $600 million in retirement benefits to deceased federal retirees over the previous five years. Checks had been illegally cashed by living relatives. One son received cumulative payments of $515,000 over the 37 years after his father died in 1971. The fraud was only discovered after the son died in 2008. (Associated Press)

sector or in the public sector. It is easy to reach agreement on opposite extremes: National defense must lie with the public sector, while automobile production can best be accomplished by the private sector. But what about health insurance? Parks and recreation areas? Fire protection? Garbage collection? Housing? Education? It is hard to assess every good or service and to say absolutely that it should be assigned to either the public sector or the private sector. Evidence: All the goods and services just mentioned are provided in part by *both* private enterprises and public agencies.

QUICK REVIEW 5.3

- Unlike the private sector—where the profit motive helps to ensure efficiency and variety—government lacks a strong incentive to be efficient and typically offers only limited and bundled choices.
- Regulatory capture occurs when a regulated industry can control its government regulator and get it to implement policies that favor the industry.
- Political corruption occurs when government officials abuse their powers for personal gain.

SUMMARY

LO5.1 Describe how government's power to coerce can be economically beneficial and list some of the difficulties associated with managing and directing the government.

Government's legal right to use coercion and force can help to improve economic efficiency by correcting for market failures and by enforcing laws and regulations that reduce the risk that individuals and firms will be taken advantage of.

LO5.2 Discuss "government failure" and explain why it happens.

Special interests can succeed in perpetuating policies that are opposed by the majority of voters because the costs of organizing and motivating groups to take political action increase with group size. This collective-action problem implies that special interests can perpetuate unpopular policies as long as the costs of organizing an opposition exceed the costs that the general public is currently suffering as a result of those policies.

There are powerful incentives for politicians to accommodate rent seeking and support special-interest legislation.

Because voters like receiving the benefits of government programs but do not like having to pay the taxes necessary to finance them, politicians tend to favor programs that offer easily identified immediate benefits but vague or deferred costs. This tendency helps to explain the unfunded liabilities of programs including Social Security as well as the federal government's tendency to run budget deficits.

When the economy goes into recession, politicians often use the need for fiscal policy stimulus as political cover to direct lower taxes or increased spending toward politically powerful special-interest groups. To prevent politicians from using lower interest rates and monetary stimulus as a way of increasing their reelection chances, most governments have put politically independent central banks in charge of monetary policy.

Economic theorists cite several reasons why government might be inefficient in providing public goods. (a) Citizens as voters and congressional representatives face limited and bundled choices as to public goods, whereas consumers in the private sector can be highly selective in their choices. (b) Government bureaucracies have less incentive to operate efficiently than do private businesses. (c) Regulated industries may sometimes capture their government regulatory agencies and mold government polices toward their own best interests.

Government's track record as an investor in private-sector firms is very poor, with most government investments into private sector businesses generating low or negative returns for taxpayers.

Government attempts to increase private investment by offering loan guarantees often cause resources to be misdirected toward high-risk projects that have an extremely low likelihood of success. These arrangements "socialize losses and privatize gains" because if the businesses go bankrupt, the government bears the losses, but if they do well, private individuals receive the profits.

Political corruption may cause governmental resources or actions to be misdirected.

Neither governments nor markets are perfect economic institutions. Each has its own set of shortcomings and citizens should be aware of where each is likely to fail and where each is likely to succeed.

TERMS AND CONCEPTS

government failure
principal-agent problems
collective-action problem
special-interest effect
earmarks
rent seeking
unfunded liability
budget deficit
debt crisis
fiscal policy
monetary policy
unintended consequences
regulatory capture
deregulation
loan guarantees
political corruption

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Why might citizens interested in maximizing economic efficiency be happy to invest their government with the right to coerce them in at least some situations? LO5.1

2. Jean-Baptiste Colbert was the Minister of Finance under King Louis XIV of France. He famously observed, "The art of taxation consists in so plucking the goose as to obtain the

largest possible amount of feathers with the smallest possible amount of hissing." How does his comment relate to special interests and the collective-action problem? **LO5.2**

3. What is rent seeking and how does it differ from the kinds of profit maximization and profit seeking that we discussed in previous chapters? Provide an actual or hypothetical example of rent seeking by firms in an industry. By a union. By a professional association (for example, physicians, school teachers, or lawyers). Why do elected officials often accommodate rent-seeking behavior, particularly by firms, unions, and professional groups located in their home states? **LO5.2**
4. How does the problem of limited and bundled choice in the public sector relate to economic efficiency? Why are public bureaucracies possibly less efficient than business firms? **LO5.2**
5. Discuss the political incentives that helped motivate federal politicians to approve budget deficits in all but four years between 1960 and 2012. **LO5.2**
6. Explain: "Politicians would make more rational economic decisions if they weren't running for reelection every few years." **LO5.2**
7. Critique: "Thank goodness we have so many government regulatory agencies. They keep Big Business in check." **LO5.2**
8. **LAST WORD** How do the concepts of pork-barrel politics and the special-interest effect relate to the items listed in the Last Word?

REVIEW QUESTIONS

1. Select all of the following that are true. To an economist, a coercive government can be useful in order to: **LO5.1**
   a. Reallocate resources in order to improve efficiency.
   b. Fight negative externalities.
   c. Ensure low gasoline prices.
   d. Provide a low-risk economic environment for individuals and firms.
2. To an economist, a government program is too big if an analysis of that program finds that MB _______ MC. **LO5.1**
   a. Is greater than.
   b. Is less than.
   c. Is equal to.
   d. Is less than twice as large as.
   e. Is more than twice as large as.
3. Tammy Hall is the mayor of a large U.S. city. She has just established the Office of Window Safety. Because windows sometimes break and spray glass shards, every window in the city will now have to pass an annual safety inspection. Property owners must pay the $5-per-window cost—and by the way, Tammy has made her nephew the new head of the Office of Window Safety. This new policy is an example of: **LO5.2**
   a. Political corruption.
   b. Earmarks.
   c. Rent seeking.
   d. Adverse selection.
4. A few hundred U.S. sugar makers lobby the U.S. government each year to make sure that the government taxes imported sugar at a high rate. They do so because the policy drives up the domestic price of sugar and increases their profits. It is estimated that the policy benefits U.S. sugar producers by about $1 billion per year while costing U.S. consumers upwards of $2 billion per year. Which of the following concepts apply to the U.S. sugar tax? **LO5.2**
   Select one or more of the choices shown.
   a. Political corruption.
   b. Rent-seeking behavior.
   c. The collective-action problem.
   d. The special-interest effect.
5. ___________________ occur when politicians commit to making a series of future expenditures without simultaneously committing to collect enough tax revenues to pay for those expenditures. **LO5.2**
   a. Budget deficits.
   b. Debt crises.
   c. Loan guarantees.
   d. Unfunded liabilities.

PROBLEMS

1. Suppose that there are 1 million federal workers at the lowest level of the federal bureaucracy and that above them there are multiple layers of supervisors and supervisors-of-supervisors. Assume that each higher level is one-tenth the size of the one below it because the government is using a 10:1 ratio of supervisees to supervisors. That is, for every 10 workers at the bottom, there is 1 supervisor; for every 10 of those supervisors, there is 1 supervisor-of-supervisors; for every one of those supervisors-of-supervisors, there is a supervisor-of-supervisors-of-supervisors; and so on, all the way up the bureaucratic pyramid to the president. **LO5.1**
   a. How many supervisors will there be in each supervisory layer of the federal bureaucracy? Start with the layer of supervisors directly above the 1 million workers at the bottom.
   b. How many supervisors are there in total at all levels of the federal bureaucratic pyramid, including the president?

c. If you count the 1 million workers at the bottom as the first layer of the federal bureaucracy, how many total layers are there, including the president?

d. How many federal employees are there in total at all layers, including the president?

e. What fraction of all federal employees are supervisory, including the president?

2. Consider a specific example of the special-interest effect and the collective-action problem. In 2009, it was estimated that the total value of all corn production subsidies in the United States was about $4 billion. The population of the United States was approximately 300 million people that year. **LO5.2**

a. On average, how much did corn subsidies cost per person in the United States in 2009? (Hint: A billion is a 1 followed by nine zeros. A million is a 1 followed by six zeros.)

b. If each person in the United States is only willing to spend $0.50 to support efforts to overturn the corn subsidy, and if antisubsidy advocates can only raise funds from 10 percent of the population, how much money will they be able to raise for their lobbying efforts?

c. If the recipients of corn subsidies donate just one percent of the total amount that they receive in subsidies, how much could they raise to support lobbying efforts to continue the corn subsidy?

d. By how many dollars does the amount raised by the recipients of the corn subsidy exceed the amount raised by the opponents of the corn subsidy?

3. Consider a corrupt provincial government in which each housing inspector examines two newly built structures each week. All the builders in the province are unethical and want to increase their profits by using substandard construction materials, but they can't do that unless they can bribe a housing inspector into approving a substandard building. **LO5.2**

a. If bribes cost $1,000 each, how much will a housing inspector make each year in bribes? (Assume that each inspector works 52 weeks a year and gets bribed for every house he inspects.)

b. There is a provincial construction supervisor who gets to hire all of the housing inspectors. He himself is corrupt and expects his housing inspectors to share their bribes with him. Suppose that 20 inspectors work for him and that each passes along half the bribes collected from builders. How much will the construction supervisor collect each year?

c. Corrupt officials may have an incentive to reduce the provision of government services to help line their own pockets. Suppose that the provincial construction supervisor decides to cut the total number of housing inspectors from 20 to 10 in order to decrease the supply of new housing permits. This decrease in the supply of permits raises the equilibrium bribe from $1,000 to $2,500. How much per year will the construction supervisor now receive if he is still getting half of all the bribes collected by the 10 inspectors? How much more is the construction supervisor getting now than when he had 20 inspectors working in part *b*? Will he personally be happy with the reduction in government services?

d. What if reducing the number of inspectors from 20 to 10 only increased the equilibrium bribe from $1,000 to $1,500? In this case, how much per year would the construction supervisor collect from his 10 inspectors? How much *less* is the construction supervisor getting than when he had 20 inspectors working in part *b*? In this case, will the construction supervisor be happy with the reduction in government services? Will he want to go back to using 20 inspectors?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

Public Choice Theory and Voting Paradoxes

LO5.3 Explain the difficulties of conveying economic preferences through majority voting.

Public Choice Theory

Market failures, such as public goods and externalities, impede economic efficiency and justify government intervention in the economy.

But the government's response to market failures is not without its own problems and pitfalls. In fact, government can sometimes fail as badly or even worse than markets in terms of delivering economic efficiency and directing resources to the uses where they will bring the largest net benefits.

That is why it is important to study **public choice theory**—the economic analysis of government decision making, politics, and elections. Just as the study of *market failure* helps us to understand how regulating markets may help to improve the allocation of resources, the study of *government failure* can help us to understand how changes in the way government functions might help it to operate more efficiently.

ORIGIN OF THE IDEA

O5.1
Public choice theory

As we will discuss shortly, many instances of government failure can be traced to incentive structures that lead political representatives to pursue policies that go against the preferences of the people that they are representing. But an even more fundamental problem exists. The majority voting systems that we rely upon may make it difficult or even impossible to correctly discern voter preferences. In such cases, it is not surprising that government fails to deliver what the voters actually want.

Revealing Preferences through Majority Voting

Through some process, society must decide which public goods it wants and in what amounts. It also must determine the extent to which it wants government to intervene in private markets to correct externalities. Decisions need to be made about the extent and type of regulation of business that is necessary, the amount of income redistribution that is desirable, what policies the government might enact to mitigate asymmetric information problems, and other such choices. Furthermore, society must determine the set of taxes it thinks is best for financing government. How should government apportion (divide) the total tax burden among the public?

Decisions such as these are made collectively in the United States through a democratic process that relies heavily on majority voting. Candidates for office offer alternative policy packages, and citizens elect people who they think will make the best decisions on their collective behalf. Voters "retire" officials who do not adequately represent their collective wishes and elect persons they think do. Also, citizens periodically have opportunities at the state and local levels to vote directly on public expenditures or new legislation.

Although the democratic process does a reasonably good job of revealing society's preferences, it is imperfect. Public choice theory demonstrates that majority voting can produce inefficiencies and inconsistencies.

Inefficient Voting Outcomes

Society's well-being is enhanced when government provides a public good whose total benefit exceeds its total cost. Unfortunately, majority voting does not always deliver that outcome.

Illustration: Inefficient "No" Vote Assume that the government can provide a public good, say, national defense, at a total expense of $900. Also assume that there are only three individuals—Adams, Benson, and Conrad—in the society and that they will share the $900 tax expense equally, each being taxed $300 if the proposed public good is provided. And assume, as Figure 1a illustrates, that Adams would receive $700 worth of benefits from having this public good; Benson, $250; and Conrad, $200.

What will be the result if a majority vote determines whether or not this public good is provided? Although people do not always vote strictly according to their own economic interest, it is likely Benson and Conrad will vote "no"

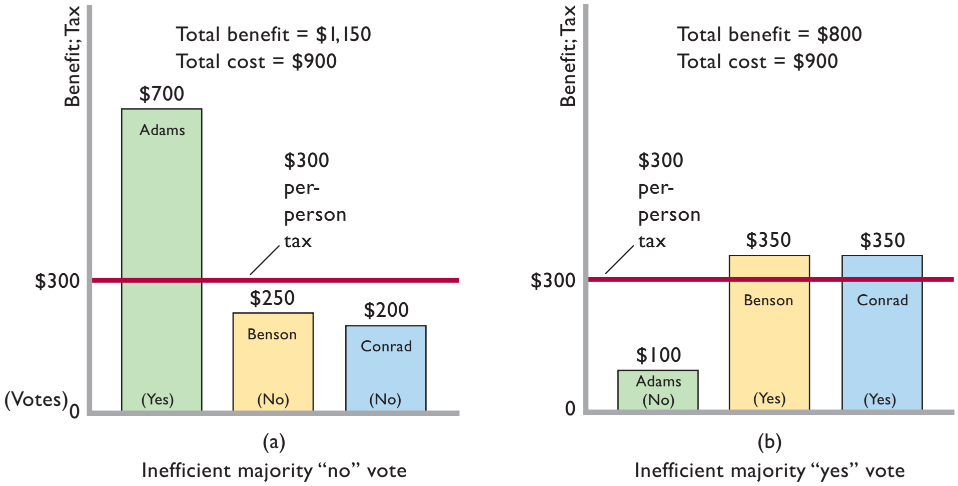


FIGURE 1 Inefficient voting outcomes. Majority voting can produce inefficient decisions. (a) Majority voting leads to rejection of a public good that would entail a greater total benefit than total cost. (b) Majority voting results in acceptance of a public good that has a higher total cost than total benefit.

because they will incur tax costs of $300 each while gaining benefits of only $250 and $200, respectively. Adams will vote "yes." So the majority vote will defeat the proposal even though the total benefit of $1,150 (= $700 for Adams + $250 for Benson + $200 for Conrad) exceeds the total cost of $900. Resources should be devoted to this good, but they will not be. Too little of this public good will be produced.

Illustration: Inefficient "Yes" Vote Now consider a situation in which the majority favors a public good even though its total cost exceeds its total benefit. Figure 1b shows the details. Again, Adams, Benson, and Conrad will equally share the $900 cost of the public good; each will be taxed $300. But since Adams' benefit now is only $100 from the public good, she will vote against it. Meanwhile, Benson and Conrad will benefit by $350 each. They will vote for the public good because that benefit ($350) exceeds their tax payments ($300). The majority vote will provide a public good costing $900 that produces total benefits of only $800 (= $100 for Adams + $350 for Benson + $350 for Conrad). Society's resources will be inefficiently allocated to this public good. Too much of it will be produced.

Implications The point is that an inefficient outcome may occur as either an overproduction or an underproduction of a specific public good, and therefore as an overallocation or underallocation of resources for that particular use. In Chapter 4 we saw that government can improve economic efficiency by providing public goods that the market system will not make available. Now we have extended that analysis to reveal that government may not provide some of those public goods or may provide them in the wrong amounts. In other cases, it may provide public goods that are not economically warranted.

In our examples, each person has only a single vote, no matter how much he or she might gain or lose from a public good. In the first example (inefficient "no" vote), Adams would be willing to purchase a vote from either Benson or Conrad if buying votes were legal. That way Adams could be assured of obtaining the national defense she so highly values. But since buying votes is illegal, many people with strong preferences for certain public goods may have to go without them.

When individual consumers have a strong preference for a specific *private good*, they usually can find that good in the marketplace even though it may be unpopular with the majority of consumers. A consumer can buy beef tongue, liver, and squid in some supermarkets, although it is doubtful that any of these products would be available if majority voting stocked the shelves. But a person cannot easily "buy" a *public good* such as national defense once the majority has decided against it.

Conversely, a consumer in the marketplace can decide against buying a particular product, even a popular one. But although you may not want national defense, you must "buy" it through your tax payments when the majority have decided they want it.

Conclusion: Because majority voting fails to incorporate the *strength* of the preferences of the individual voter, it may produce economically inefficient outcomes.

Interest Groups and Logrolling Some, but not all, of the inefficiencies of majority voting get resolved through the political process. Two examples follow.

Interest Groups People who share strong preferences for a public good may band together into interest groups and use advertisements, mailings, and direct persuasion to convince others of the merits of that public good. Adams might try to persuade Benson and Conrad that it is in their best interest to vote for national defense—that national defense is much more valuable to them than their $250 and $200 valuations. Such appeals are common in democratic politics. Sometimes they are successful; sometimes they are not.

Political Logrolling Perhaps surprisingly, **logrolling**—the trading of votes to secure desired outcomes—can also turn an inefficient outcome into an efficient one. In our first example (Figure 1a), suppose that Benson has a strong preference for a different public good, for example, a new road, which Adams and Conrad do not think is worth the tax expense. That would provide an opportunity for Adams and Benson to trade votes to ensure provision of both national defense and the new road. That is, Adams and Benson would each vote "yes" on both measures. Adams would get the national defense and Benson would get the road. Without the logrolling, both public goods would have been rejected. This logrolling will add to society's well-being if, as was true for national defense, the road creates a greater overall benefit than cost.

But logrolling need not increase economic efficiency. Even if national defense and the road each cost more than the total benefit each produces, both might still be provided if there is vote trading. Adams and Benson might still engage in logrolling if each expects to secure a sufficient net gain from her or his favored public good, even though the gains would come at the clear expense of Conrad.

Logrolling is very common in state legislatures and Congress. It can either increase or diminish economic efficiency, depending on the circumstances.

Paradox of Voting

ORIGIN OF THE IDEA

O5.2

Paradox of voting

Another difficulty with majority voting is the **paradox of voting,** a situation in which society may not be able to rank its preferences consistently through paired-choice majority voting.

Preferences Consider Table 1, in which we again assume a community of three voters: Adams, Benson, and Conrad. Suppose the community has three alternative public goods from which to choose: national defense, a road, and a weather warning system. We expect that each member of the community prefers the three alternatives in a certain order. For example, one person might prefer national defense to a road and a road to a weather warning system. We can attempt to determine the preferences of the community through paired-choice majority voting. Specifically, a vote can be held between any two of the public goods, and the winner of that vote can then be matched against the third public good in another vote.

The three goods and the assumed individual preferences of the three voters are listed in the top part of Table 1.

TABLE 1 Paradox of Voting

	Preferences		
Public Good	**Adams**	**Benson**	**Conrad**
National defense	1st choice	3d choice	2d choice
Road	2d choice	1st choice	3d choice
Weather warning system	3d choice	2d choice	1st choice
Election	**Voting Outcomes: Winner**		
1. National defense vs. road	National defense (preferred by Adams and Conrad)		
2. Road vs. weather warning system	Road (preferred by Adams and Benson)		
3. National defense vs. weather warning system	Weather warning system (preferred by Benson and Conrad)		

The data indicate that Adams prefers national defense to the road and the road to the weather warning system. This implies also that Adams prefers national defense to the weather warning system. Benson values the road more than the weather warning system and the warning system more than national defense. Conrad's order of preference is weather warning system, national defense, and road.

Voting Outcomes The lower part of Table 1 shows the outcomes of three hypothetical elections decided through majority vote. In the first, national defense wins against the road because a majority of voters (Adams and Conrad) prefer national defense to the road. In the second election, to see whether this community wants a road or a weather warning system, a majority of voters (Adams and Benson) prefer the road.

We have determined that the majority of people in this community prefer national defense to a road and prefer a road to a weather warning system. It seems logical to conclude that the community prefers national defense to a weather warning system. But it does not!

To demonstrate this conclusion, we hold a direct election between national defense and the weather warning system. Row 3 shows that a majority of voters (Benson and Conrad) prefer the weather warning system to national defense. As listed in Table 1, then, the three paired-choice majority votes imply that this community is irrational: It seems to prefer national defense to a road and a road to a weather warning system, but would rather have a weather warning system than national defense.

The problem is not irrational community preferences but rather a flawed procedure for determining those preferences. We see that the outcome from paired-choice majority voting may depend on the order in which the votes are taken. Different sequences of majority votes can lead to different outcomes, many of which may fail to reflect the electorate's underlying preferences. As a consequence, government may find it difficult to provide the "correct" public goods by acting in accordance with majority voting. Important note: This critique is not meant to suggest that some better procedure exists. Majority voting is much more likely to reflect community preferences than decisions by, say, a dictator or a group of self-appointed leaders.

Median-Voter Model

One other aspect of majority voting reveals further insights into real-world phenomena. The **median-voter model** suggests that, under majority rule and consistent voting preferences, the median voter will in a sense determine the outcomes of elections. The median voter is the person holding the middle position on an issue: Half the other voters have stronger preferences for a public good, amount of taxation, or degree of government regulation, while half have weaker or negative preferences. The extreme voters on each side of an issue prefer the median choice rather than the other extreme position, so the median voter's choice predominates.

Example Suppose a society composed of Adams, Benson, and Conrad has reached agreement that as a society it needs a weather warning system. Each person independently is to submit a total dollar amount he or she thinks should be spent on the warning system, assuming each will be taxed one-third of that amount. An election will determine the size of the system. Because each person can be expected to vote for his or her own proposal, no majority will occur if all the proposals are placed on the ballot at the same time. Thus, the group decides on a paired-choice vote: They will first vote between two of the proposals and then match the winner of that vote against the remaining proposal.

The three proposals are as follows: Adams desires a $400 system; Benson wants an $800 system; Conrad opts for a $300 system. Which proposal will win? The median-voter model suggests it will be the $400 proposal submitted by the median voter, Adams. Half the other voters favor a more costly system; half favor a less costly system. To understand why the $400 system will be the outcome, let's conduct the two elections.

First, suppose that the $400 proposal is matched against the $800 proposal. Adams naturally votes for her $400 proposal, and Benson votes for his own $800 proposal. Conrad, who proposed the $300 expenditure for the warning system, votes for the $400 proposal because it is closer to his own. So Adams' $400 proposal is selected by a 2-to-1 majority vote.

Next, we match the $400 proposal against the $300 proposal. Again the $400 proposal wins. It gets a vote from Adams and one from Benson, who proposed the $800 expenditure and for that reason prefers a $400 expenditure to a $300 one. Adams, the median voter in this case, is in a sense the person who has decided the level of expenditure on a weather warning system for this society.

Real-World Applicability Although our illustration is simple, it explains a great deal. We do note a tendency for public choices to match most closely the median view. Political candidates, for example, take one set of positions to win the nomination of their political parties; in so doing, they tend to appeal to the median voter within the party to get the nomination. They then shift their views more closely to the political center when they square off against opponents from the opposite political party. In effect, they redirect their appeal toward

CONSIDER THIS . . .

Voter Failure

Inefficient voting outcomes and the paradox of voting imply that governments may sometimes fail to deliver the best combination of public goods because it may be very difficult for politicians to discern what voters actually want. In other cases, though, economists worry that governments may end up failing to deliver allocative and productive efficiency not because politicians can't tell what people want—but because they *can*.

The problem is that voters sometimes support policies that reduce rather than enhance allocative and productive efficiency. Examples include several types of wage and price controls, punitive tariffs on foreign products, and various industrial and agricultural subsidies.

These policies almost always reduce economic efficiency, but they are also extremely popular with voters in many countries. Faced with that reality, a politician may well end up supporting such policies even if he personally understands that they will create more economic harm than benefit.

That behavior makes some observers wish for braver politicians who might be willing to oppose these instances of "voter failure." But others argue that it is too much to hope for braver politicians. Instead, efforts should be directed toward educating the public and convincing them to support government policies that are economically efficient.

the median voter within the total population. They also try to label their opponents as being too liberal, or too conservative, and out of touch with "mainstream America." And they conduct polls and adjust their positions on issues accordingly.

Implications The median-voter model has two important implications:

- At any point in time, many people will be dissatisfied by the extent of government involvement in the economy. The size of government will largely be determined by the median preference, leaving many people desiring a much larger, or a much smaller, public sector. In the marketplace you can buy no zucchinis, 2 zucchinis, or 200 zucchinis, depending on how much you enjoy them. In the public sector you will tend to get the number of Stealth bombers and new highway projects that the median voter prefers.
- Some people may "vote with their feet" by moving into political jurisdictions where the median voter's preferences are closer to their own. They may move from the city to a suburb where the level of government services, and therefore taxes, is lower. Or they may move into an area known for its excellent, but expensive, school system. Some may move to other states; a few may even move to other countries.

For these reasons, and because our personal preferences for publicly provided goods and services are not static, the median preference shifts over time. Moreover, information about people's preferences is imperfect, leaving much room for politicians to misjudge the true median position. When they do, they may have a difficult time getting elected or reelected.

APPENDIX SUMMARY

LO5.3 Explain the difficulties of conveying economic preferences through majority voting.

Public choice theory suggests that governments may sometimes suffer from government failures because majority voting fails to correctly indicate voter preferences.

Majority voting creates the possibility of (a) underallocations or overallocations of resources to particular public goods and (b) inconsistent voting outcomes that make it impossible for a democratic political system to definitively determine the will of the people.

The median-voter model predicts that, under majority rule, the person holding the middle position on an issue will determine the outcome of an election involving that issue.

APPENDIX TERMS AND CONCEPTS

public choice theory
paradox of voting
median-voter model
logrolling

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Explain how affirmative and negative majority votes can sometimes lead to inefficient allocations of resources to public goods. Is this problem likely to be greater under a benefits-received or an ability-to-pay tax system? Use the information in Figures 1a and 1b to show how society might be better off if Adams were allowed to buy votes. **LO5.3**
2. "Majority voting ensures that government will produce only those public goods for which benefits exceed costs." Discuss. **LO5.3**
3. "The problem with our democratic institutions is that they don't correctly reflect the will of the people! If the people—rather than self-interested politicians or lobbyists—had control, we wouldn't have to worry about government taking actions that don't maximize allocative and productive efficiency." Critique. **LO5.3**

APPENDIX REVIEW QUESTIONS

1. Explain the paradox of voting through reference to the accompanying table, which shows the ranking of three public goods by voters Jay, Dave, and Conan: **LO5.3**

	Rankings		
Public Good	**Jay**	**Dave**	**Conan**
Courthouse	2nd choice	1st choice	3d choice
School	3d choice	2d choice	1st choice
Park	1st choice	3d choice	2d choice

2. We can apply voting paradoxes to the highway construction example of Chapter 4. Suppose there are only five people in a society and each favors one of the five highway construction options listed in Table 4.4 ("No new construction" is one of the five options). Explain which of these highway options will be selected using a majority paired-choice vote. Will this option be the optimal size of the project from an economic perspective? **LO5.3**
3. True or False: The median-voter model explains why politicians so often stake out fringe positions that appeal only to a small segment of the electorate. **LO5.3**

APPENDIX PROBLEMS

1. Look back at Figures 1a and 1b, which show the costs and benefits to voters Adams, Benson, and Conrad of two different public goods that the government will produce if a majority of Adams, Benson, and Conrad support them. Suppose that Adams, Benson, and Conrad have decided to have one single vote at which the funding for both of those public goods will be decided simultaneously. **LO5.3**
   - **a.** Given the $300 cost per person of each public good, what are Adams' net benefits for each public good individually and for the two combined? Will he want to vote yes or no on the proposal to fund both projects simultaneously?
   - **b.** What are Conrad's net benefits for each public good individually and for the two combined? Will he want to vote yes or no on the proposal to fund both projects simultaneously?
   - **c.** What are Benson's net benefits for each public good individually and for the two combined? Will he want to vote yes or no on the proposal to fund both projects simultaneously—or will he be indifferent?
   - **d.** Who is the median voter here? Who will the two other voters be attempting to persuade?
2. Political advertising is often directed at winning over so-called swing voters, whose votes might go either way. Suppose that two political parties—the Freedom Party and the Liberty Party—disagree on whether to build a new road. Polling shows that of 1,000 total voters, 450 are firmly for the new road and 450 are firmly against the new road. Thus, each party will try to win over a majority of the 100 remaining swing voters. **LO5.3**
   - **a.** Suppose that each party spends $5,000 on untargeted TV, radio, and newspaper ads that are equally likely to reach any and all voters. How much per voter will be spent by both parties combined?
   - **b.** Suppose that, instead, each party could direct all of its spending toward just the swing voters by using targeted ads that exploit Internet social media. If all of the two parties' combined spending was targeted at just swing voters, how much would be spent per swing voter?
   - **c.** Suppose that only the Freedom Party knows how to target voters using social media. How much per swing voter will it be spending? If at the same time the Liberty Party is still using only untargeted TV, radio, and newspaper ads, what portion of its total spending is likely to be reaching the 100 swing voters? How much per swing voter does that portion amount to?
   - **d.** Looking at your answers to part *c*, how much more per swing voter will the Freedom Party be spending than the Liberty Party? If spending per swing voter influences elections, which party is more likely to win?

PART **THREE**

CONSUMER BEHAVIOR

CHAPTER **6**

Elasticity

Learning Objectives:

LO6.1 Discuss price elasticity of demand and how it is calculated.

LO6.2 Explain the usefulness of the total revenue test for price elasticity of demand.

LO6.3 List the factors that affect price elasticity of demand and describe some applications of price elasticity of demand.

LO6.4 Describe price elasticity of supply and how it can be applied.

LO6.5 Apply cross elasticity of demand and income elasticity of demand.

In this chapter we extend Chapter 3's discussion of demand and supply by explaining *elasticity,* an extremely important concept that helps us answer such questions as: Why do buyers of some products (for example, ocean cruises) respond to price increases by substantially reducing their purchases while buyers of other products (say, gasoline) respond by only slightly cutting back their purchases? Why do higher market prices for some products (for example, chicken) cause producers to greatly increase their output while price rises for other products (say, gold) cause only limited increases in output? Why does the demand for some products (for example, books) rise a great deal when household income increases while the demand for other products (say, milk) rises just a little?

Elasticity extends our understanding of markets by letting us know the degree to which changes in prices and incomes affect supply and demand. Sometimes the responses are substantial, other times minimal or even nonexistent. But by knowing what to expect, businesses and the government can do a better job in deciding what to produce, how much to charge, and, surprisingly, what items to tax.

Price Elasticity of Demand

LO6.1 Discuss price elasticity of demand and how it is calculated.

The law of demand tells us that, other things equal, consumers will buy more of a product when its price declines and less when its price increases. But how much more or less will they buy? The amount varies from product to product and over different price ranges for the same product. It also may vary over time. And such variations matter. For example, a firm contemplating a price hike will want to know how consumers will respond. If they remain highly loyal and continue to buy, the firm's revenue will rise. But if consumers defect en masse to other sellers or other products, the firm's revenue will tumble.

The responsiveness (or sensitivity) of consumers to a price change is measured by a product's **price elasticity of demand.** For some products—for example, restaurant meals—consumers are highly responsive to price changes. Modest price changes cause very large changes in the quantity purchased. Economists say that the demand for such products is *relatively elastic* or simply *elastic*.

ORIGIN OF THE IDEA

O6.1
Price elasticity of demand

For other products—for example, toothpaste—consumers pay much less attention to price changes. Substantial price changes cause only small changes in the amount purchased. The demand for such products is *relatively inelastic* or simply *inelastic*.

The Price-Elasticity Coefficient and Formula

Economists measure the degree to which demand is price elastic or inelastic with the coefficient E_d, defined as

$$E_d = \frac{\text{percentage change in quantity demanded of product X}}{\text{percentage change in price of product X}}$$

The percentage changes in the equation are calculated by dividing the *change* in quantity demanded by the original quantity demanded and by dividing the *change* in price by the original price. So we can restate the formula as

$$E_d = \frac{\text{change in quantity demanded of X}}{\text{original quantity demanded of X}} \div \frac{\text{change in price of X}}{\text{original price of X}}$$

Using Averages Unfortunately, an annoying problem arises in computing the price-elasticity coefficient. A price change from, say, \$4 to \$5 along a demand curve is a 25 percent (= \$1/\$4) increase, but the opposite price change from \$5 to \$4 along the same curve is a 20 percent (= \$1/\$5) decrease. Which percentage change in price should we use in the denominator to compute the price-elasticity coefficient? And when quantity changes, for example, from 10 to 20, it is a 100 percent (=10/10) increase. But when quantity falls from 20 to 10 along the identical demand curve, it is a 50 percent (=10/20) decrease. Should we use 100 percent or 50 percent in the numerator of the elasticity formula? Elasticity should be the same whether price rises or falls!

The simplest solution to the problem is to use the **midpoint formula** for calculating elasticity. This formula simply averages the two prices and the two quantities as the reference points for computing the percentages. That is,

$$E_d = \frac{\text{change in quantity}}{\text{sum of quantities}/2} \div \frac{\text{change in price}}{\text{sum of prices}/2}$$

For the same \$5–\$4 price range, the price reference is \$4.50 [= (\$5 + \$4)/2], and for the same 10–20 quantity range, the quantity reference is 15 units [= (10 + 20)/2]. The percentage change in price is now \$1/\$4.50, or about 22 percent, and the percentage change in quantity is $\frac{10}{15}$, or about 67 percent. So E_d is about 3. This solution eliminates the "up versus down" problem. All the price-elasticity coefficients that follow are calculated using this midpoint formula.

WORKED PROBLEMS

W6.1
Elasticity of demand

Using Percentages Why use percentages rather than absolute amounts in measuring consumer responsiveness? There are two reasons.

First, if we use absolute changes, the choice of units will arbitrarily affect our impression of buyer responsiveness. To illustrate: If the price of a bag of popcorn at the local softball game is reduced from \$3 to \$2 and consumers increase their purchases from 60 to 100 bags, it will seem that consumers are quite sensitive to price changes and therefore that demand is elastic. After all, a price change of 1 unit has caused a change in the amount demanded of 40 units. But by changing the monetary unit from dollars to pennies (why not?), we find that a price change of 100 units (pennies) causes a quantity change of 40 units. This may falsely lead us to believe that demand is inelastic. We avoid this problem by using percentage changes. This particular price decline is the same whether we measure it in dollars or pennies.

Second, by using percentages, we can correctly compare consumer responsiveness to changes in the prices of different products. It makes little sense to compare the effects on quantity demanded of (1) a $1 increase in the price of a $10,000 used car with (2) a $1 increase in the price of a $1 soft drink. Here the price of the used car has increased by 0.01 percent while the price of the soft drink is up by 100 percent. We can more sensibly compare the consumer responsiveness to price increases by using some common percentage increase in price for both.

Elimination of Minus Sign We know from the downsloping demand curve that price and quantity demanded are inversely related. Thus, the price-elasticity coefficient of demand E_d will always be a negative number. As an example, if price declines, then quantity demanded will increase. This means that the numerator in our formula will be positive and the denominator negative, yielding a negative E_d. For an increase in price, the numerator will be negative but the denominator positive, again yielding a negative E_d.

Economists usually ignore the minus sign and simply present the absolute value of the elasticity coefficient to avoid an ambiguity that might otherwise arise. It can be confusing to say that an E_d of -4 is greater than one of -2. This possible confusion is avoided when we say an E_d of 4 reveals greater elasticity than one of 2. So, in what follows, we ignore the minus sign in the coefficient of price elasticity of demand and show only the absolute value. Incidentally, the ambiguity does not arise with supply because price and quantity supplied are positively related. All elasticity of supply coefficients therefore are positive numbers.

Interpretations of E_d

We can interpret the coefficient of price elasticity of demand as follows.

Elastic Demand Demand is **elastic** if a specific percentage change in price results in a larger percentage change in quantity demanded. In such cases, E_d will be greater than 1. Example: Suppose that a 2 percent decline in the price of cut flowers results in a 4 percent increase in quantity demanded. Then demand for cut flowers is elastic and

$$E_d = \frac{.04}{.02} = 2$$

Inelastic Demand If a specific percentage change in price produces a smaller percentage change in quantity demanded, demand is **inelastic.** In such cases, E_d will be less than 1. Example: Suppose that a 2 percent decline in the price of coffee leads to only a 1 percent increase in quantity demanded. Then demand is inelastic and

$$E_d = \frac{.01}{.02} = .5$$

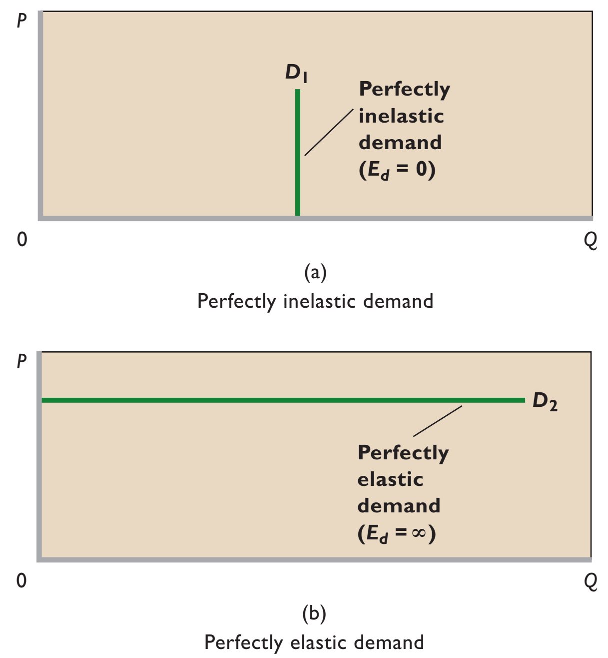
Unit Elasticity The case separating elastic and inelastic demands occurs where a percentage change in price and the resulting percentage change in quantity demanded are the same. Example: Suppose that a 2 percent drop in the price of chocolate causes a 2 percent increase in quantity demanded. This special case is termed **unit elasticity** because E_d is exactly 1, or unity. In this example,

$$E_d = \frac{.02}{.02} = 1$$

Extreme Cases When we say demand is "inelastic," we do not mean that consumers are completely unresponsive to a price change. In that extreme situation, where a price change results in no change whatsoever in the quantity demanded, economists say that demand is **perfectly inelastic.** The price-elasticity coefficient is zero because there is no response to a change in price. Approximate examples include an acute diabetic's demand for insulin or an addict's demand for heroin. A line parallel to the vertical axis, such as D_1 in Figure 6.1a, shows perfectly inelastic demand graphically.

Conversely, when we say demand is "elastic," we do not mean that consumers are completely responsive to a price change. In that extreme situation, where a small price reduction causes buyers to increase their purchases from zero to all they can obtain, the elasticity coefficient is infinite ($= \infty$) and economists say demand is **perfectly elastic.** A line parallel to the horizontal axis, such as D_2 in Figure 6.1b, shows perfectly elastic demand. You will see in Chapter 10 that such a demand applies to a firm—say, a mining firm—that is selling its output in a purely competitive market.

FIGURE 6.1 Perfectly inelastic and elastic demands. Demand curve D_1 in (a) represents perfectly inelastic demand ($E_d = 0$). A price increase will result in no change in quantity demanded. Demand curve D_2 in (b) represents perfectly elastic demand. A price increase will cause quantity demanded to decline from an infinite amount to zero ($E_d = \infty$).



The Total-Revenue Test

LO6.2 Explain the usefulness of the total revenue test for price elasticity of demand.

The importance of elasticity for firms relates to the effect of price changes on total revenue and thus on profits (= total revenue minus total costs).

Total revenue (TR) is the total amount the seller receives from the sale of a product in a particular time period; it is calculated by multiplying the product price (P) by the quantity sold (Q). In equation form:

$$\text{TR} = P \times Q$$

Graphically, total revenue is represented by the $P \times Q$ rectangle lying below a point on a demand curve. At point *a* in Figure 6.2a, for example, price is \$2 and quantity demanded is 10 units. So total revenue is \$20 (= \$2 × 10), shown by the rectangle composed of the yellow and green areas under the demand curve. We know from basic geometry that the area of a rectangle is found by multiplying one side by the other. Here, one side is "price" (\$2) and the other is "quantity demanded" (10 units).

Total revenue and the price elasticity of demand are related. In fact, the easiest way to infer whether demand is elastic or inelastic is to employ the **total-revenue test.** Here is the test: Note what happens to total revenue when price changes. If total revenue changes in the opposite direction from price, demand is elastic. If total revenue changes in the same direction as price, demand is inelastic. If total revenue does not change when price changes, demand is unit-elastic.

Elastic Demand

If demand is elastic, a decrease in price will increase total revenue. Even though a lesser price is received per unit, enough additional units are sold to more than make up for the lower price. For an example, look at demand curve D_1

CONSIDER THIS . . .

A Bit of a Stretch

The following analogy might help you remember the distinction between "elastic" and "inelastic." Imagine two objects—one an Ace elastic bandage used to wrap injured joints and the other a relatively firm rubber tie-down (rubber strap) used for securing items for transport. The Ace bandage stretches a great deal when pulled with a particular force; the rubber tie-down stretches some, but not a lot.

Similar differences occur for the quantity demanded of various products when their prices change. For some products, a price change causes a substantial "stretch" of quantity demanded. When this stretch in percentage terms exceeds the percentage change in price, demand is elastic. For other products, quantity demanded stretches very little in response to the price change. When this stretch in percentage terms is less than the percentage change in price, demand is inelastic.

In summary:

- Elastic demand displays considerable "quantity stretch" (as with the Ace bandage).
- Inelastic demand displays relatively little "quantity stretch" (as with the rubber tie-down).

And through extension:

- Perfectly elastic demand has infinite quantity stretch.
- Perfectly inelastic demand has zero quantity stretch.

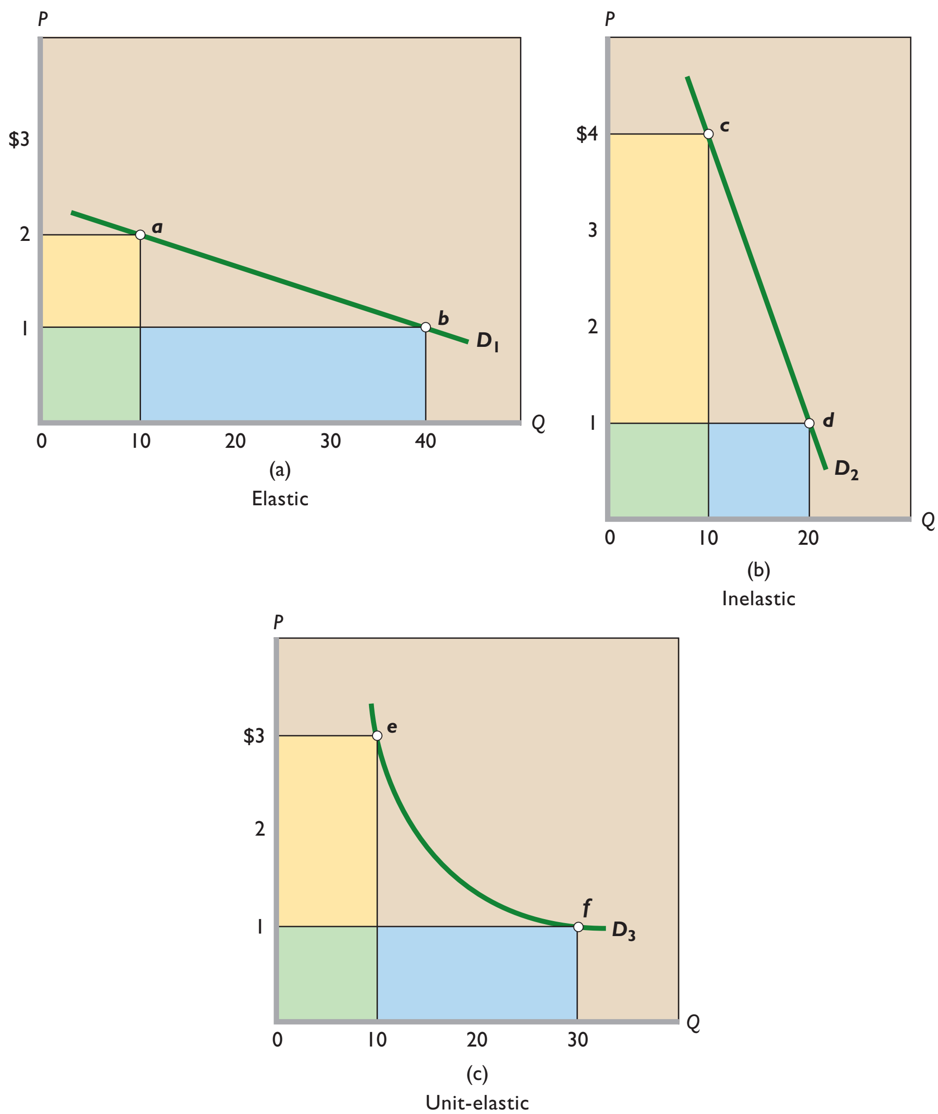


FIGURE 6.2 The total-revenue test for price elasticity. (a) Price declines from \$2 to \$1, and total revenue increases from \$20 to \$40. So demand is elastic. The gain in revenue (blue area) exceeds the loss of revenue (yellow area). (b) Price declines from \$4 to \$1, and total revenue falls from \$40 to \$20. So, demand is inelastic. The gain in revenue (blue area) is less than the loss of revenue (yellow area). (c) Price declines from \$3 to \$1, and total revenue does not change. Demand is unit-elastic. The gain in revenue (blue area) equals the loss of revenue (yellow area).

in Figure 6.2a. We have already established that at point *a*, total revenue is \$20 (= \$2 × 10), shown as the yellow plus green area. If the price declines from \$2 to \$1 (point *b*), the quantity demanded becomes 40 units and total revenue is \$40 (= \$1 × 40). As a result of the price decline, total revenue has increased from \$20 to \$40. Total revenue has increased in this case because the \$1 decline in price applies to 10 units, with a consequent revenue loss of \$10 (the yellow area). But 30 more units are sold at \$1 each, resulting in a revenue gain of \$30 (the blue area). Visually, the gain of the blue area clearly exceeds the loss of the yellow area. As indicated, the overall result is a net increase in total revenue of \$20 (= \$30 − \$10).

The analysis is reversible: If demand is elastic, a price increase will reduce total revenue. The revenue gained on the higher-priced units will be more than offset by the revenue lost from the lower quantity sold. Bottom line: Other things equal, when price and total revenue move in opposite directions, demand is elastic. E_d is greater than 1, meaning the percentage change in quantity demanded is greater than the percentage change in price.

Inelastic Demand

If demand is inelastic, a price decrease will reduce total revenue. The increase in sales will not fully offset the decline in revenue per unit, and total revenue will decline. To see this, look at demand curve D_2 in Figure 6.2b. At point *c* on the curve, price is \$4 and quantity demanded is 10. Thus total revenue is \$40, shown by the combined yellow and green rectangle. If the price drops to $1 (point *d*), total revenue declines to $20, which obviously is less than $40. Total revenue has declined because the loss of revenue (the yellow area) from the lower unit price is larger than the gain in revenue (the blue area) from the accompanying increase in sales. Price has fallen, and total revenue has also declined.

Our analysis is again reversible: If demand is inelastic, a price increase will increase total revenue. So, other things equal, when price and total revenue move in the same direction, demand is inelastic. E_d is less than 1, meaning the percentage change in quantity demanded is less than the percentage change in price.

WORKED PROBLEMS

W6.2
Total-revenue test

Unit Elasticity

In the special case of unit elasticity, an increase or a decrease in price leaves total revenue unchanged. The loss in revenue from a lower unit price is exactly offset by the gain in revenue from the accompanying increase in sales. Conversely, the gain in revenue from a higher unit price is exactly offset by the revenue loss associated with the accompanying decline in the amount demanded.

In Figure 6.2c (demand curve D_3) we find that at the price of $3, 10 units will be sold, yielding total revenue of $30. At the lower $1 price, a total of 30 units will be sold, again resulting in $30 of total revenue. The $2 price reduction causes the loss of revenue shown by the yellow area, but this is exactly offset by the revenue gain shown by the blue area. Total revenue does not change. In fact, that would be true for all price changes along this particular curve.

Other things equal, when price changes and total revenue remains constant, demand is unit-elastic (or unitary). E_d is 1, meaning the percentage change in quantity equals the percentage change in price.

Price Elasticity along a Linear Demand Curve

Now a major confession! Although the demand curves depicted in Figure 6.2 nicely illustrate the total-revenue test for elasticity, two of the graphs involve specific movements along linear (straight-line) demand curves. That presents no problem for explaining the total-revenue test. However, you need to know that elasticity typically varies over different price ranges of the same demand curve. (The exception is the curve in Figure 6.2c. Elasticity is 1 along the entire curve.)

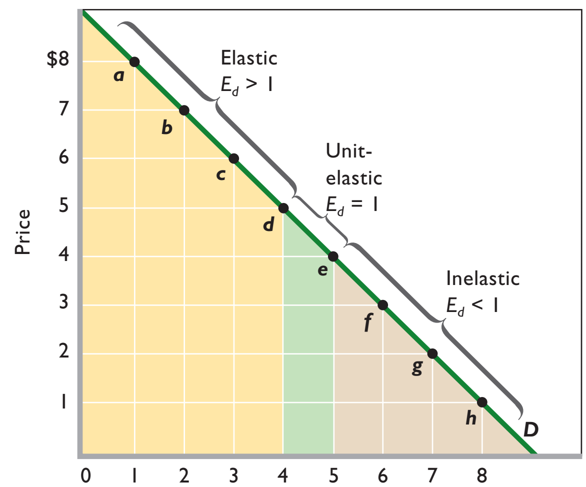
Table 6.1 and Figure 6.3 demonstrate that elasticity typically varies over different price ranges of the same demand schedule or curve. Plotting the hypothetical data for movie tickets shown in columns 1 and 2 of Table 6.1 yields demand curve *D* in Figure 6.3. Observe that the demand curve is linear. But we see from column 3 of the table that the price elasticity coefficient for this demand curve declines as we move from higher to lower prices. For all downsloping straight-line and most other demand curves, demand is more price-elastic toward the upper left (here, the $5–$8 price range of *D*) than toward the lower right (here, the $4–$1 price range of *D*).

This is the consequence of the arithmetic properties of the elasticity measure. Specifically, in the upper-left segment of the demand curve, the percentage change in quantity is large because the original reference quantity is small. Similarly, the percentage change in price is small in that segment because the original reference price is large. The relatively large percentage change in quantity divided

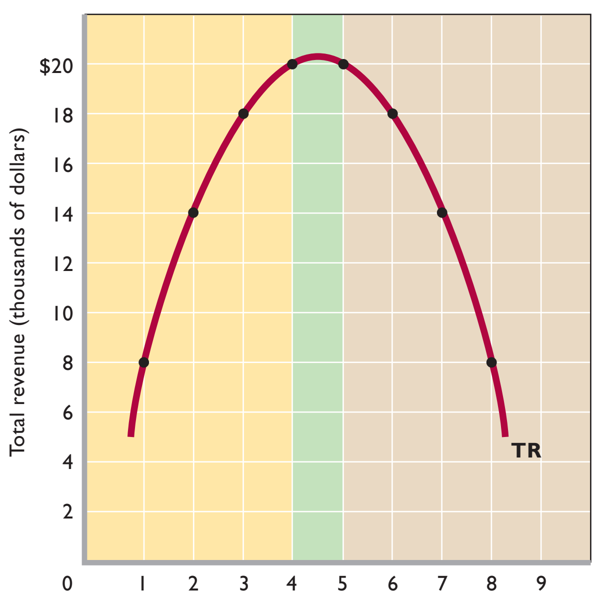
TABLE 6.1 **Price Elasticity of Demand for Movie Tickets as Measured by the Elasticity Coefficient and the Total-Revenue Test**

(1) Total Quantity of Tickets Demanded per Week, Thousands	(2) Price per Ticket	(3) Elasticity Coefficient (E_d)	(4) Total Revenue, (1) × (2)	(5) Total-Revenue Test
1	$8		$ 8,000	
		5.00		Elastic
2	7		14,000	
		2.60		Elastic
3	6		18,000	
		1.57		Elastic
4	5		20,000	
		1.00		Unit-elastic
5	4		20,000	
		0.64		Inelastic
6	3		18,000	
		0.38		Inelastic
7	2		14,000	
		0.20		Inelastic
8	1		8,000	

FIGURE 6.3 The relation between price elasticity of demand for movie tickets and total revenue. (a) Demand curve *D* is based on Table 6.1 and is marked to show that the hypothetical weekly demand for movie tickets is elastic at higher price ranges and inelastic at lower price ranges. (b) The total-revenue curve TR is derived from demand curve *D*. When price falls and TR increases, demand is elastic; when price falls and TR is unchanged, demand is unit-elastic; and when price falls and TR declines, demand is inelastic.



(a)
Demand curve



(b)
Total-revenue curve

by the relatively small change in price yields a large E_d—an elastic demand.

The reverse holds true for the lower-right segment of the demand curve. Here the percentage change in quantity is small because the original reference quantity is large; similarly, the percentage change in price is large because the original reference price is small. The relatively small percentage change in quantity divided by the relatively large percentage change in price results in a small E_d—an inelastic demand.

The demand curve in Figure 6.3a also illustrates that the slope of a demand curve—its flatness or steepness—is not a sound basis for judging elasticity. The catch is that the slope of the curve is computed from *absolute* changes in price and quantity, while elasticity involves *relative* or *percentage* changes in price and quantity. The demand curve in Figure 6.3a is linear, which by definition means that the slope is constant throughout. But we have demonstrated that such a curve is elastic in its high-price ($8–$5) range and inelastic in its low-price ($4–$1) range.

Price Elasticity and the Total-Revenue Curve

In Figure 6.3b we plot the total revenue per week to the theater owner that corresponds to each price-quantity combination indicated along demand curve *D* in Figure 6.3a. The price–quantity-demanded combination represented by point *a* on the demand curve yields total revenue of $8,000 (= $8 × 1,000 tickets). In Figure 6.3b, we plot this $8,000 amount vertically at 1 unit (1,000 tickets) demanded. Similarly, the price–quantity-demanded combination represented by point *b* in the upper panel yields total revenue of $14,000 (= $7 × 2,000 tickets). This amount is graphed vertically at 2 units (2,000 tickets) demanded in the lower panel. The ultimate result of such graphing is total-revenue curve TR, which first slopes upward, then reaches a maximum, and finally turns downward.

Comparison of curves *D* and TR sharply focuses the relationship between elasticity and total revenue. Lowering the ticket price in the elastic range of demand—for example, from $8 to $5—increases total revenue. Conversely, increasing the ticket price in that range reduces total revenue. In both cases, price and total revenue change in opposite directions, confirming that demand is elastic.

The $5–$4 price range of demand curve *D* reflects unit elasticity. When price either decreases from $5 to $4 or increases from $4 to $5, total revenue remains $20,000. In both cases, price has changed and total revenue has remained constant, confirming that demand is unit-elastic when we consider these particular price changes.

In the inelastic range of demand curve *D*, lowering the price—for example, from $4 to $1—decreases total revenue, as shown in Figure 6.3b. Raising the price boosts

TABLE 6.2 **Price Elasticity of Demand: A Summary**

Absolute Value of Elasticity Coefficient	Demand Is:	Description	Impact on Total Revenue of a: Price Increase	Price Decrease
Greater than 1 ($E_d > 1$)	Elastic or relatively elastic	Quantity demanded changes by a larger percentage than does price	Total revenue decreases	Total revenue increases
Equal to 1 ($E_d = 1$)	Unit- or unitary elastic	Quantity demanded changes by the same percentage as does price	Total revenue is unchanged	Total revenue is unchanged
Less than 1 ($E_d < 1$)	Inelastic or relatively inelastic	Quantity demanded changes by a smaller percentage than does price	Total revenue increases	Total revenue decreases

total revenue. In both cases, price and total revenue move in the same direction, confirming that demand is inelastic.

Table 6.2 summarizes the characteristics of price elasticity of demand. You should review it carefully.

Determinants of Price Elasticity of Demand

LO6.3 List the factors that affect price elasticity of demand and describe some applications of price elasticity of demand.

We cannot say just what will determine the price elasticity of demand in each individual situation. However, the following generalizations are often helpful.

- ***Substitutability*** Generally, the larger the number of substitute goods that are available, the greater the price elasticity of demand. Various brands of candy bars are generally substitutable for one another, making the demand for one brand of candy bar, say Snickers, highly elastic. Toward the other extreme, the demand for tooth repair (or tooth pulling) is quite inelastic because there simply are no close substitutes when those procedures are required.

  The elasticity of demand for a product depends on how narrowly the product is defined. Demand for Reebok sneakers is more elastic than is the overall demand for shoes. Many other brands are readily substitutable for Reebok sneakers, but there are few, if any, good substitutes for shoes.
- ***Proportion of Income*** Other things equal, the higher the price of a good relative to consumers' incomes, the greater the price elasticity of demand. A 10 percent increase in the price of low-priced pencils or chewing gum amounts to a few more pennies relative to a consumer's income, and quantity demanded will probably decline only slightly. Thus, price elasticity for such low-priced items tends to be low. But a 10 percent increase in the price of relatively high-priced automobiles or housing means additional expenditures of perhaps \$3,000 or \$20,000, respectively. These price increases are significant fractions of the annual incomes and budgets of most families, and quantities demanded will likely diminish significantly. The price elasticities for such items tend to be high.
- ***Luxuries versus Necessities*** In general, the more that a good is considered to be a "luxury" rather than a "necessity," the greater is the price elasticity of demand. Electricity is generally regarded as a necessity; it is difficult to get along without it. A price increase will not significantly reduce the amount of lighting and power used in a household. (Note the very low price-elasticity coefficient of this good in Table 6.3.) An extreme case: A person does not decline an operation for acute appendicitis because the physician's fee has just gone up.

  On the other hand, vacation travel and jewelry are luxuries, which, by definition, can easily be forgone. If the prices of vacation travel and jewelry rise, a consumer need not buy them and will suffer no great hardship without them.

  What about the demand for a common product like salt? It is highly inelastic on three counts: Few good substitutes are available; salt is a negligible item in the family budget; and it is a "necessity" rather than a luxury.
- ***Time*** Generally, product demand is more elastic the longer the time period under consideration. Consumers often need time to adjust to changes in prices. For example, when the price of a product rises, time is needed to find and experiment with other products to see if they are acceptable. Consumers may not immediately reduce their purchases very much when the price of beef rises by 10 percent, but in time they may shift to chicken, pork, or fish.

Another consideration is product durability. Studies show that "short-run" demand for gasoline is more inelastic

TABLE 6.3 **Selected Price Elasticities of Demand**

Product or Service	Coefficient of Price Elasticity of Demand (E_d)	Product or Service	Coefficient of Price Elasticity of Demand (E_d)
Newspapers	.10	Milk	.63
Electricity (household)	.13	Household appliances	.63
Bread	.15	Liquor	.70
Major League Baseball tickets	.23	Movies	.87
Cigarettes	.25	Beer	.90
Telephone service	.26	Shoes	.91
Sugar	.30	Motor vehicles	1.14
Medical care	.31	Beef	1.27
Eggs	.32	China, glassware, tableware	1.54
Legal services	.37	Residential land	1.60
Automobile repair	.40	Restaurant meals	2.27
Clothing	.49	Lamb and mutton	2.65
Gasoline	.60	Fresh peas	2.83

Source: Compiled from numerous studies and sources reporting price elasticity of demand.

($E_d = 0.2$) than is "long-run" demand ($E_d = 0.7$). In the short run, people are "stuck" with their present cars and trucks, but with rising gasoline prices they eventually replace them with smaller, more fuel-efficient vehicles. They also switch to mass transit where it is available.

Table 6.3 shows estimated price-elasticity coefficients for a number of products. Each reflects some combination of the elasticity determinants just discussed.

Applications of Price Elasticity of Demand

The concept of price elasticity of demand has great practical significance, as the following examples suggest.

Large Crop Yields The demand for most farm products is highly inelastic; E_d is perhaps 0.20 or 0.25. As a result, increases in the supply of farm products arising from a good growing season or from increased productivity tend to depress both the prices of farm products and the total revenues (incomes) of farmers. For farmers as a group, the inelastic demand for their products means that large crop yields may be undesirable. For policymakers it means that achieving the goal of higher total farm income requires that farm output be restricted.

Excise Taxes The government pays attention to elasticity of demand when it selects goods and services on which to levy excise taxes. If a \$1 tax is levied on a product and 10,000 units are sold, tax revenue will be \$10,000 (= \$1 × 10,000 units sold). If the government raises the tax to \$1.50, but the higher price that results reduces sales to 4,000 because of elastic demand, tax revenue will decline to \$6,000 (= \$1.50 × 4,000 units sold). Because a higher tax on a product with elastic demand will bring in less tax revenue, legislatures tend to seek out products that have inelastic demand—such as liquor, gasoline, and cigarettes—when levying excises.

Decriminalization of Illegal Drugs In recent years proposals to legalize drugs have been widely debated. Proponents contend that drugs should be treated like alcohol; they should be made legal for adults and regulated for purity and potency. The current war on drugs, it is argued, has been unsuccessful, and the associated costs—including enlarged police forces, the construction of more prisons, an overburdened court system, and untold human costs—have increased markedly. Legalization would allegedly reduce drug trafficking significantly by taking the profit out of it. Crack cocaine and heroin, for example, are cheap to produce and could be sold at low prices in legal markets. Because the demand of addicts is highly inelastic, the amounts consumed at the lower prices would increase only modestly. Addicts' total expenditures for cocaine and heroin would decline, and so would the street crime that finances those expenditures.

Opponents of legalization say that the overall demand for cocaine and heroin is far more elastic than proponents think. In addition to the inelastic demand of addicts, there is another market segment whose demand is relatively elastic. This segment consists of the occasional users or "dabblers," who use hard drugs when their prices are low but who abstain or substitute, say, alcohol when their prices are high. Thus, the lower prices associated with the legalization of hard drugs would increase consumption by dabblers. Also, removal of the legal prohibitions against using drugs might make drug use more socially acceptable, increasing the demand for cocaine and heroin.

Many economists predict that the legalization of cocaine and heroin would reduce street prices by up to 60 percent, depending on if and how much they were taxed. According to an important study, price declines of that size would increase the number of occasional users of heroin by 54 percent and the number of occasional users of cocaine by 33 percent. The total quantity of heroin demanded would rise by an estimated 100 percent, and the quantity of cocaine demanded would rise by 50 percent.[1] Moreover, many existing and first-time dabblers might in time become addicts. The overall result, say the opponents of legalization, would be higher social costs, possibly including an increase in street crime.

QUICK REVIEW 6.1

- The price elasticity of demand coefficient E_d is the ratio of the percentage change in quantity demanded to the percentage change in price. The *averages* of the two prices and two quantities are used as the base references in calculating the percentage changes.
- When E_d is greater than 1, demand is elastic; when E_d is less than 1, demand is inelastic; when E_d is equal to 1, demand is of unit elasticity.
- When price changes, total revenue will change in the opposite direction if demand is price-elastic, in the same direction if demand is price-inelastic, and not at all if demand is unit-elastic.
- Demand is typically elastic in the high-price (low-quantity) range of the demand curve and inelastic in the low-price (high-quantity) range of the demand curve.
- Price elasticity of demand is greater (a) the larger the number of substitutes available; (b) the higher the price of a product relative to one's budget; (c) the greater the extent to which the product is a luxury; and (d) the longer the time period involved.

Price Elasticity of Supply

LO6.4 Describe price elasticity of supply and how it can be applied.

ORIGIN OF THE IDEA

O6.2

Price elasticity of supply

The concept of price elasticity also applies to supply. If the quantity supplied by producers is relatively responsive to price changes, supply is elastic. If it is relatively insensitive to price changes, supply is inelastic.

We measure the degree of price elasticity or inelasticity of supply with the coefficient E_s, defined almost like E_d except that we substitute "percentage change in quantity supplied" for "percentage change in quantity demanded":

$$E_s = \frac{\text{percentage change in quantity supplied of product X}}{\text{percentage change in price of product X}}$$

For reasons explained earlier, the averages, or midpoints, of the before and after quantities supplied and the before and after prices are used as reference points for the percentage changes. Suppose an increase in the price of a good from \$4 to \$6 increases the quantity supplied from 10 units to 14 units. The percentage change in price would be $\frac{2}{5}$, or 40 percent, and the percentage change in quantity would be $\frac{4}{12}$, or 33 percent. Consequently,

$$E_s = \frac{.33}{.40} = .83$$

In this case, supply is inelastic because the price-elasticity coefficient is less than 1. If E_s is greater than 1, supply is elastic. If it is equal to 1, supply is unit-elastic. Also, E_s is never negative, since price and quantity supplied are directly related. Thus, there are no minus signs to drop, as was necessary with elasticity of demand.

The degree of **price elasticity of supply** depends on how easily—and therefore quickly—producers can shift resources between alternative uses. The easier and more rapidly producers can shift resources between alternative uses, the greater the price elasticity of supply. Take the case of Christmas trees. A firm's response to, say, an increase in the price of trees depends on its ability to shift resources from the production of other products (whose prices we assume remain constant) to the production of trees. And shifting resource takes time: The longer the time, the greater the "shiftability." So we can expect a greater response, and therefore greater elasticity of supply, the longer a firm has to adjust to a price change.

[1]Henry Saffer and Frank Chaloupka, "The Demand for Illegal Drugs," *Economic Inquiry*, July 1999, pp. 401–411.

CONSIDER THIS . . .

Elasticity and College Costs

Why does college cost so much? Elasticity offers some clues.

From the end of World War II through the 1970s, the supply of higher education increased massively as state and local governments spent billions of dollars expanding their higher education systems. This massive increase in supply helped to offset the huge increase in demand that took place as the large Baby Boom generation flooded the higher education system starting in the early 1960s. With supply increasing nearly as fast as demand, the equilibrium price of higher education only increased modestly.

Things changed dramatically beginning in the early 1980s. With respect to supply, state and local governments slowed the growth of higher education spending, so that many college and university systems saw only modest subsequent increases in capacity. At the same time, the federal government dramatically increased both subsidized student lending and the volume of federal student grant money. Those policy innovations were of great benefit to poor and middle-class students, but they also meant that the demand curve for higher education continued to shift to the right.

That turned out to be problematic because, with the supply of seats largely fixed by the changing priorities of state and local governments, the supply of higher education was highly inelastic even in the long run. As a result, the increases in demand caused by student loans and grant money resulted in substantially higher equilibrium prices for higher education.

In response, some economists propose that the best way to increase affordability and access would be to put more priority on the pre-1980 policies that increased supply rather than demand.

In analyzing the impact of time on elasticity, economists distinguish among the immediate market period, the short run, and the long run.

Price Elasticity of Supply: The Immediate Market Period

The **immediate market period** is the length of time over which producers are unable to respond to a change in price with a change in quantity supplied. Suppose the owner of a small farm brings to market one truckload of tomatoes that is the entire season's output. The supply curve for the tomatoes is perfectly inelastic (vertical); the farmer will sell the truckload whether the price is high or low. Why? Because the farmer can offer only one truckload of tomatoes even if the price of tomatoes is much higher than anticipated. The farmer might like to offer more tomatoes, but tomatoes cannot be produced overnight. Another full growing season is needed to respond to a higher-than-expected price by producing more than one truckload. Similarly, because the product is perishable, the farmer cannot withhold it from the market. If the price is lower than anticipated, the farmer will still sell the entire truckload.

The farmer's costs of production, incidentally, will not enter into this decision to sell. Though the price of tomatoes may fall far short of production costs, the farmer will nevertheless sell everything he brought to market to avoid a total loss through spoilage. In the immediate market period, both the supply of tomatoes and the quantity of tomatoes supplied are fixed. The farmer offers only one truckload no matter how high or low the price.

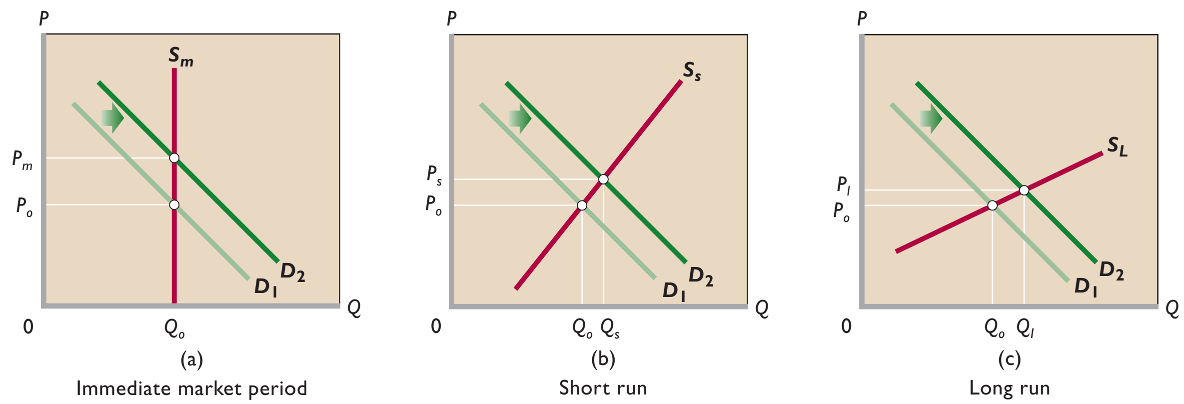
Figure 6.4a shows the farmer's vertical supply curve during the immediate market period. Supply is perfectly inelastic because the farmer does not have time to respond to a change in demand, say, from D_1 to D_2. The resulting price increase from P_0 to P_m simply determines which buyers get the fixed quantity supplied; it elicits no increase in output.

However, not all supply curves are perfectly inelastic immediately after a price change. If the product is not perishable and the price rises, producers may choose to increase quantity supplied by drawing down their inventories of unsold, stored goods. This will cause the market supply curve to attain some positive slope. For our tomato farmer, the immediate market period may be a full growing season; for producers of goods that can be inexpensively stored, there may be no immediate market period at all.

Price Elasticity of Supply: The Short Run

The **short run** in microeconomics is a period of time too short to change plant capacity but long enough to use the fixed-sized plant more or less intensively. In the short run, our farmer's plant (land and farm machinery) is fixed. But he does have time in the short run to cultivate tomatoes more intensively by applying more labor and more fertilizer and pesticides to the crop. The result is a somewhat greater output in response to a presumed increase in

FIGURE 6.4 Time and the elasticity of supply. The greater the amount of time producers have to adjust to a change in demand, here from D_1 to D_2, the greater will be their output response. (a) In the immediate market period, there is insufficient time to change output, and so supply is perfectly inelastic. (b) In the short run, plant capacity is fixed, but changing the intensity of its use can alter output; supply is therefore more elastic. (c) In the long run, all desired adjustments, including changes in plant capacity, can be made, and supply becomes still more elastic.



demand; this greater output is reflected in a more elastic supply of tomatoes, as shown by S_s in Figure 6.4b. Note now that the increase in demand from D_1 to D_2 is met by an increase in quantity (from Q_0 to Q_s), so there is a smaller price adjustment (from P_0 to P_s) than would be the case in the immediate market period. The equilibrium price is therefore lower in the short run than in the immediate market period.

Price Elasticity of Supply: The Long Run

The **long run** in microeconomics is a time period long enough for firms to adjust their plant sizes and for new firms to enter (or existing firms to leave) the industry. In the "tomato industry," for example, our farmer has time to acquire additional land and buy more machinery and equipment. Furthermore, other farmers may, over time, be attracted to tomato farming by the increased demand and higher price. Such adjustments create a larger supply response, as represented by the more elastic supply curve S_L in Figure 6.4c. The outcome is a smaller price rise (P_0 to P_l) and a larger output increase (Q_0 to Q_l) in response to the increase in demand from D_1 to D_2.

There is no total-revenue test for elasticity of supply. Supply shows a positive or direct relationship between price and amount supplied; the supply curve is upsloping. Regardless of the degree of elasticity or inelasticity, price and total revenue always move together.

Applications of Price Elasticity of Supply

The idea of price elasticity of supply has widespread applicability, as suggested by the following examples.

Antiques and Reproductions *Antiques Roadshow* is a popular PBS television program in which people bring antiques to a central location for appraisal by experts. Some people are pleased to learn that their old piece of furniture or funky folk art is worth a large amount, say, $30,000 or more.

The high price of an antique results from strong demand and limited, highly inelastic supply. Because a genuine antique can no longer be reproduced, its quantity supplied either does not rise or rises only slightly as price goes up. The higher price might prompt the discovery of a few more of the remaining originals and thus add to the quantity available for sale, but this quantity response is usually quite small. So the supply of antiques and other collectibles tends to be inelastic. For one-of-a-kind antiques, the supply is perfectly inelastic.

Factors such as increased population, higher income, and greater enthusiasm for collecting antiques have increased the demand for antiques over time. Because the supply of antiques is limited and inelastic, those increases in demand have greatly boosted the prices of antiques.

Contrast the inelastic supply of original antiques with the elastic supply of modern "made-to-look-old" reproductions. Such faux antiques are quite popular and widely available at furniture stores and knickknack shops. When the demand for reproductions increases, the firms making them simply boost production. Because the supply of reproductions is highly elastic, increased demand raises their prices only slightly.

Volatile Gold Prices The price of gold is quite volatile, sometimes shooting upward one period and plummeting

LAST WORD

Elasticity and Pricing Power: Why Different Consumers Pay Different Prices

Firms and Nonprofit Institutions Often Recognize and Exploit Differences in Price Elasticity of Demand.

All the buyers of a product traded in a highly competitive market pay the same market price for the product, regardless of their individual price elasticities of demand. If the price rises, Jones may have an elastic demand and greatly reduce her purchases. Green may have a unit-elastic demand and reduce his purchases less than Jones. Lopez may have an inelastic demand and hardly curtail his purchases at all. But all three consumers will pay the single higher price regardless of their respective demand elasticities.

In later chapters we will find that not all sellers must passively accept a "one-for-all" price. Some firms have "market power" or "pricing power" that allows them to set their product prices in their best interests. For some goods and services, firms may find it advantageous to determine differences in price elasticity of demand and then charge different prices to different buyers.

It is extremely difficult to tailor prices for each customer on the basis of price elasticity of demand, but it is relatively easy to observe differences in group elasticities. Consider airline tickets. Business travelers generally have inelastic demand for air travel. Because their time is highly valuable, they do not see slower modes of transportation as realistic substitutes. Also, their employers pay for their tickets as part of their business expenses. In contrast, leisure travelers tend to have elastic demand. They have the option to drive rather than fly or to simply not travel at all. They also pay for their tickets out of their own pockets and thus are more sensitive to price.

Airlines recognize the difference between the groups in terms of price elasticity of demand and charge business travelers more than leisure travelers. To accomplish that, they have to dissuade business travelers from buying the less expensive round-trip tickets aimed at leisure travelers. One way to do this is by placing restrictions on the lower-priced tickets. For instance, airlines have at times made such tickets nonrefundable, required at least a 2-week advance purchase, and required Saturday-night stays. These restrictions chase off most business travelers who engage in last-minute travel and want to be home for the weekend. As a result, a business traveler often pays hundreds of dollars more for a ticket than a leisure traveler on the same plane.

downward the next. The main sources of these fluctuations are shifts in demand interacting with highly inelastic supply. Gold production is a costly and time-consuming process of exploration, mining, and refining. Moreover, the physical availability of gold is highly limited. For both reasons, increases in gold prices do not elicit substantial increases in quantity supplied. Conversely, gold mining is costly to shut down and existing gold bars are expensive to store. Price decreases therefore do not produce large drops in the quantity of gold supplied. In short, the supply of gold is inelastic.

The demand for gold is partly derived from the demand for its uses, such as for jewelry, dental fillings, and coins. But people also demand gold as a speculative financial investment. They increase their demand for gold when they fear general inflation or domestic or international turmoil that might undermine the value of currency and more traditional investments. They reduce their demand when events settle down. Because of the inelastic supply of gold, even relatively small changes in demand produce relatively large changes in price. (This chapter's Web-based question 1 that is posted online provides an Internet source for finding current and past prices of gold.)

Cross Elasticity and Income Elasticity of Demand

LO6.5 Apply cross elasticity of demand and income elasticity of demand.

Price elasticities measure the responsiveness of the quantity of a product demanded or supplied when its price changes. The consumption of a good also is affected by a change in the price of a related product or by a change in income.

Cross Elasticity of Demand

The **cross elasticity of demand** measures how sensitive consumer purchases of one product (say, X) are to a change

Discounts for children are another example of pricing based on group differences in price elasticity of demand. For many products, children have more elastic demands than adults because children have low budgets, often financed by their parents. Sellers recognize the elasticity difference and price accordingly. The barber spends as much time cutting a child's hair as an adult's but charges the child much less. A child takes up a full seat at the baseball game but pays a lower price than an adult. A child snowboarder occupies the same space on a chairlift as an adult snowboarder but qualifies for a discounted lift ticket.

Finally, consider pricing by colleges and universities. Price elasticity of demand for higher education is greater for prospective students from low-income families than similar students from high-income families. This makes sense because tuition is a much larger proportion of household income for a low-income student or family than for his or her high-income counterpart. Desiring a diverse student body, colleges charge different *net* prices (= tuition *minus* financial aid) to the two groups on the basis of price elasticity of demand. High-income students pay full tuition, unless they receive merit-based scholarships. Low-income students receive considerable financial aid in addition to merit-based scholarships and, in effect, pay a lower *net* price.

It is common for colleges to announce a large tuition increase and immediately cushion the news by emphasizing that they also are increasing financial aid. In effect, the college is increasing the tuition for students who have inelastic demand by the full amount and raising the *net* tuition of those with elastic demand by some lesser amount or not at all. Through this strategy, colleges boost revenue to cover rising costs while maintaining affordability for a wide range of students.

There are a number of other examples of dual or multiple pricing. All relate directly to price elasticity of demand. We will revisit this topic again in Chapter 12 when we analyze *price discrimination*—charging different prices to different customers for the same product.

in the price of some other product (say, Y). We calculate the coefficient of cross elasticity of demand E_{xy} just as we do the coefficient of simple price elasticity, except that we relate the percentage change in the consumption of X to the percentage change in the price of Y:

$$E_{xy} = \frac{\text{percentage change in quantity demanded of product X}}{\text{percentage change in price of product Y}}$$

This cross-elasticity (or cross-price-elasticity) concept allows us to quantify and more fully understand substitute and complementary goods, introduced in Chapter 3. Unlike price elasticity, we allow the coefficient of cross elasticity of demand to be either positive or negative.

Substitute Goods If cross elasticity of demand is positive, meaning that sales of X move in the same direction as a change in the price of Y, then X and Y are substitute goods. An example is Evian water (X) and Dasani water (Y). An increase in the price of Evian causes consumers to buy more Dasani, resulting in a positive cross elasticity. The larger the positive cross-elasticity coefficient, the greater is the substitutability between the two products.

Complementary Goods When cross elasticity is negative, we know that X and Y "go together"; an increase in the price of one decreases the demand for the other. So the two are complementary goods. For example, a decrease in the price of digital cameras will increase the number of memory sticks purchased. The larger the negative cross-elasticity coefficient, the greater is the complementarity between the two goods.

Independent Goods A zero or near-zero cross elasticity suggests that the two products being considered are unrelated or independent goods. An example is walnuts and plums: We would not expect a change in the price of walnuts to have any effect on purchases of plums, and vice versa.

Application The degree of substitutability of products, measured by the cross-elasticity coefficient, is important to businesses and government. For example, suppose that Coca-Cola is considering whether or not to lower the price of its Sprite brand. Not only will it want to know something about the price elasticity of demand for Sprite (will the price cut increase or decrease total revenue?), but it will also be interested in knowing if the increased sales of Sprite will come at the expense of its Coke brand. How sensitive are the sales of one of its products (Coke) to a change in the price of another of its products (Sprite)? By how much will the increased sales of Sprite "cannibalize" the sales of Coke? A low cross elasticity would indicate that Coke and Sprite are weak substitutes for each other and that a lower price for Sprite would have little effect on Coke sales.

Government also implicitly uses the idea of cross elasticity of demand in assessing whether a proposed merger between two large firms will substantially reduce competition and therefore violate the antitrust laws. For example, the cross elasticity between Coke and Pepsi is high, making them strong substitutes for each other. In addition, Coke and Pepsi together sell about 70 percent of all carbonated cola drinks consumed in the United States. Taken together, the high cross elasticities and the large market shares suggest that the government would likely block a merger between Coke and Pepsi because the merger would substantially lessen competition. In contrast, the cross elasticity between cola and gasoline is low or zero. A merger between Coke and Shell oil company would have a minimal effect on competition. So government would let that merger happen.

Income Elasticity of Demand

Income elasticity of demand measures the degree to which consumers respond to a change in their incomes by buying more or less of a particular good. The coefficient of income elasticity of demand E_i is determined with the formula

$$E_i = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in income}}$$

Normal Goods For most goods, the income-elasticity coefficient E_i is positive, meaning that more of them are demanded as incomes rise. Such goods are called normal or superior goods (and were first described in Chapter 3). But the value of E_i varies greatly among normal goods. For example, income elasticity of demand for automobiles is about $+3$, while income elasticity for most farm products is only about $+0.20$.

Inferior Goods A negative income-elasticity coefficient designates an inferior good. Retread tires, cabbage, long-distance bus tickets, used clothing, and muscatel wine are likely candidates. Consumers decrease their purchases of inferior goods as incomes rise.

Insights Coefficients of income elasticity of demand provide insights into the economy. For example, when recessions (business downturns) occur and incomes fall, income elasticity of demand helps predict which products will decline in demand more rapidly than others.

Products with relatively high income elasticity coefficients, such as automobiles ($E_i = +3$), housing ($E_i = +1.5$), and restaurant meals ($E_i = +1.4$), are generally hit hardest by recessions. Those with low or negative income elasticity coefficients are much less affected. For example, food products prepared at home ($E_i = +0.20$) respond relatively little to income fluctuations. When incomes drop, purchases of food (and toothpaste and toilet paper) drop little compared to purchases of movie tickets, luxury vacations, and plasma screen TVs. Products we view as essential tend to have lower income elasticity coefficients than products we view as luxuries. When our incomes fall, we cannot easily eliminate or postpone the purchase of essential products.

In Table 6.4 we provide a convenient synopsis of the cross-elasticity and income-elasticity concepts.

QUICK REVIEW 6.2

- Price elasticity of supply measures the sensitivity of suppliers to changes in the price of a product. The price-elasticity-of-supply coefficient E_s is the ratio of the percentage change in quantity supplied to the percentage change in price. The elasticity of supply varies directly with the amount of time producers have to respond to the price change.
- The cross-elasticity-of-demand coefficient E_{xy} is computed as the percentage change in the quantity demanded of product X divided by the percentage change in the price of product Y. If the cross-elasticity coefficient is positive, the two products are substitutes; if negative, they are complements.
- The income-elasticity coefficient E_i is computed as the percentage change in quantity demanded divided by the percentage change in income. A positive coefficient indicates a normal or superior good. The coefficient is negative for an inferior good.

TABLE 6.4 **Cross and Income Elasticities of Demand**

Value of Coefficient	Description	Type of Good(s)
Cross elasticity:		
Positive ($E_{wz} > 0$)	Quantity demanded of W changes in same direction as change in price of Z	Substitutes
Negative ($E_{xy} < 0$)	Quantity demanded of X changes in opposite direction from change in price of Y	Complements
Income elasticity:		
Positive ($E_i > 0$)	Quantity demanded of the product changes in same direction as change in income	Normal or superior
Negative ($E_i < 0$)	Quantity demanded of the product changes in opposite direction from change in income	Inferior

SUMMARY

LO6.1 Discuss price elasticity of demand and how it is calculated.

Price elasticity of demand measures consumer response to price changes. If consumers are relatively sensitive to price changes, demand is elastic. If they are relatively unresponsive to price changes, demand is inelastic.

The price-elasticity coefficient E_d measures the degree of elasticity or inelasticity of demand. The coefficient is found by the formula

$$E_d = \frac{\text{percentage change in quantity demanded of X}}{\text{percentage change in price of X}}$$

Economists use the averages of prices and quantities under consideration as reference points in determining percentage changes in price and quantity. If E_d is greater than 1, demand is elastic. If E_d is less than 1, demand is inelastic. Unit elasticity is the special case in which E_d equals 1.

Perfectly inelastic demand is graphed as a line parallel to the vertical axis; perfectly elastic demand is shown by a line above and parallel to the horizontal axis.

Elasticity varies at different price ranges on a demand curve, tending to be elastic in the upper-left segment and inelastic in the lower-right segment. Elasticity cannot be judged by the steepness or flatness of a demand curve.

LO6.2 Explain the usefulness of the total revenue test for price elasticity of demand.

If total revenue changes in the opposite direction from prices, demand is elastic. If price and total revenue change in the same direction, demand is inelastic. Where demand is of unit elasticity, a change in price leaves total revenue unchanged.

LO6.3 List the factors that affect price elasticity of demand and describe some applications of price elasticity of demand.

The number of available substitutes, the size of an item's price relative to one's budget, whether the product is a luxury or a necessity, and length of time to adjust are all determinants of elasticity of demand.

LO6.4 Describe price elasticity of supply and how it can be applied.

The elasticity concept also applies to supply. The coefficient of price elasticity of supply is found by the formula

$$E_s = \frac{\text{percentage change in quantity supplied of X}}{\text{percentage change in price of X}}$$

The averages of the prices and quantities under consideration are used as reference points for computing percentage changes. Elasticity of supply depends on the ease of shifting resources between alternative uses, which varies directly with the time producers have to adjust to a price change.

LO6.5 Apply cross elasticity of demand and income elasticity of demand.

Cross elasticity of demand indicates how sensitive the purchase of one product is to changes in the price of another product. The coefficient of cross elasticity of demand is found by the formula

$$E_{xy} = \frac{\text{percentage change in quantity demanded of X}}{\text{percentage change in price of Y}}$$

Positive cross elasticity of demand identifies substitute goods; negative cross elasticity identifies complementary goods.

Income elasticity of demand indicates the responsiveness of consumer purchases to a change in income. The coefficient of income elasticity of demand is found by the formula

$$E_i = \frac{\text{percentage change in quantity demanded of X}}{\text{percentage change in income}}$$

The coefficient is positive for normal goods and negative for inferior goods.

Industries that sell products that have high income elasticity of demand coefficients are particularly hard hit by recessions. Those with products that have low or negative income elasticity of demand coefficients fare much better.

TERMS AND CONCEPTS

price elasticity of demand
midpoint formula
elastic demand
inelastic demand
unit elasticity
perfectly inelastic demand
perfectly elastic demand
total revenue (TR)
total-revenue test
price elasticity of supply
immediate market period
short run
long run
cross elasticity of demand
income elasticity of demand

DISCUSSION QUESTIONS

1. Explain why the choice between 1, 2, 3, 4, 5, 6, 7, and 8 "units," or 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, and 8,000 movie tickets, makes no difference in determining elasticity in Table 6.1. **LO6.1**
2. What effect would a rule stating that university students must live in university dormitories have on the price elasticity of demand for dormitory space? What impact might this in turn have on room rates? **LO6.1**
3. The income elasticities of demand for movies, dental services, and clothing have been estimated to be +3.4, +1, and + 0.5, respectively. Interpret these coefficients. What does it mean if an income elasticity coefficient is negative? **LO6.5**
4. Research has found that an increase in the price of beer would reduce the amount of marijuana consumed. Is cross elasticity of demand between the two products positive or negative? Are these products substitutes or complements? What might be the logic behind this relationship? **LO6.5**
5. **LAST WORD** What is the purpose of charging different groups of customers different prices? Supplement the three broad examples in the Last Word with two additional examples of your own. Hint: Think of price discounts based on group characteristics or time of purchase.

REVIEW QUESTIONS

1. Suppose that the total revenue received by a company selling basketballs is $600 when the price is set at $30 per basketball and $600 when the price is set at $20 per basketball. Without using the midpoint formula, can you tell whether demand is elastic, inelastic, or unit-elastic over this price range? **LO6.2**
2. What are the major determinants of price elasticity of demand? Use those determinants and your own reasoning in judging whether demand for each of the following products is probably elastic or inelastic: (*a*) bottled water; (*b*) toothpaste, (*c*) Crest toothpaste, (*d*) ketchup, (*e*) diamond bracelets, (*f*) Microsoft's Windows operating system. **LO6.3**
3. Calculate total-revenue data from the demand schedule in review question 1. Graph total revenue below your demand curve. Generalize about the relationship between price elasticity and total revenue. **LO6.2**
4. How would the following changes in price affect total revenue? That is, would total revenue increase, decrease, or remain unchanged? **LO6.2**
   - **a.** Price falls and demand is inelastic.
   - **b.** Price rises and demand is elastic.
   - **c.** Price rises and supply is elastic.
   - **d.** Price rises and supply is inelastic.
   - **e.** Price rises and demand is inelastic.
   - **f.** Price falls and demand is elastic.
   - **g.** Price falls and demand is of unit elasticity.
5. In 2006, Willem de Kooning's abstract painting *Woman III* sold for $137.5 million. Portray this sale in a demand and supply diagram and comment on the elasticity of supply. Comedian George Carlin once mused, "If a painting can be forged well enough to fool some experts, why is the original so valuable?" Provide an answer. **LO6.4**
6. Suppose the cross elasticity of demand for products A and B is +3.6 and for products C and D is −5.4. What can you conclude about how products A and B are related? Products C and D? **LO6.5**

PROBLEMS

1. Look at the demand curve in Figure 6.2a. Use the midpoint formula and points *a* and *b* to calculate the elasticity of demand for that range of the demand curve. Do the same for the demand curves in Figures 6.2b and 6.2c using, respectively, points *c* and *d* for Figure 6.2b and points *e* and *f* for Figure 6.2c. **LO6.1**
2. Investigate how demand elastiticities are affected by increases in demand. Shift each of the demand curves in Figures 6.2a, 6.2b, and 6.2c to the right by 10 units. For example, point *a* in Figure 6.2a would shift rightward from location (10 units, $2) to (20 units, $2), while point *b* would shift rightward from location (40 units, $1) to (50 units, $1). After making these shifts, apply the midpoint formula to calculate the demand elasticities for the shifted points. Are they larger or smaller than the elasticities you calculated in problem 1 for the original points? In terms of the midpoint formula, what explains the change in elasticities? **LO6.1**
3. Graph the accompanying demand data, and then use the midpoint formula for E_d to determine price elasticity of demand for each of the four possible $1 price changes. What can you conclude about the relationship between the slope of a curve and its elasticity? Explain in a nontechnical way why demand is elastic in the northwest segment of the demand curve and inelastic in the southeast segment. **LO6.1**

Product Price	Quantity Demanded
$5	1
4	2
3	3
2	4
1	5

4. Danny "Dimes" Donahue is a neighborhood's 9-year-old entrepreneur. His most recent venture is selling homemade brownies that he bakes himself. At a price of $1.50 each, he sells 100. At a price of $1 each, he sells 300. Is demand elastic or inelastic over this price range? If demand had the same elasticity for a price decline from $1.00 to $0.50 as it does for the decline from $1.50 to $1, would cutting the price from $1.00 to $0.50 increase or decrease Danny's total revenue? **LO6.2**
5. What is the formula for measuring the price elasticity of supply? Suppose the price of apples goes up from $20 to $22 a box. In direct response, Goldsboro Farms supplies 1,200 boxes of apples instead of 1,000 boxes. Compute the coefficient of price elasticity (midpoints approach) for Goldsboro's supply. Is its supply elastic, or is it inelastic? **LO6.4**
6. **ADVANCED ANALYSIS** Currently, at a price of $1 each, 100 popsicles are sold per day in the perpetually hot town of Rostin. Consider the elasticity of supply. In the short run, a price increase from $1 to $2 is unit-elastic ($E_s = 1.0$). So how many popsicles will be sold each day in the short run if the price rises to $2 each? In the long run, a price increase from $1 to $2 has an elasticity of supply of 1.50. So how many popsicles will be sold per day in the long run if the price rises to $2 each? (Hint: Apply the midpoints approach to the elasticity of supply.) **LO6.4**
7. Lorena likes to play golf. The number of times per year that she plays depends on both the price of playing a round of golf as well as Lorena's income and the cost of other types of entertainment—in particular, how much it costs to go see a movie instead of playing golf. The three demand schedules in the table below show how many rounds of golf per year Lorena will demand at each price under three different scenarios. In scenario D_1, Lorena's income is $50,000 per year and movies cost $9 each. In scenario D_2, Lorena's income is also $50,000 per year, but the price of seeing a movie rises to $11. And in scenario D_3, Lorena's income goes up to $70,000 per year, while movies cost $11. **LO6.5**

Price	Quantity Demanded D_1	D_2	D_3
$50	15	10	15
35	25	15	30
20	40	20	50

   a. Using the data under D_1 and D_2, calculate the cross elasticity of Lorena's demand for golf at all three prices. (To do this, apply the midpoints approach to the cross elasticity of demand.) Is the cross elasticity the same at all three prices? Are movies and golf substitute goods, complementary goods, or independent goods?

   b. Using the data under D_2 and D_3, calculate the income elasticity of Lorena's demand for golf at all three prices. (To do this, apply the midpoints approach to the income elasticity of demand.) Is the income elasticity the same at all three prices? Is golf an inferior good?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER 7

Utility Maximization

Learning Objectives

LO7.1 Define and explain the relationship between total utility, marginal utility, and the law of diminishing marginal utility.

LO7.2 Describe how rational consumers maximize utility by comparing the marginal utility-to-price ratios of all the products they could possibly purchase.

LO7.3 Explain how a demand curve can be derived by observing the outcomes of price changes in the utility-maximization model.

LO7.4 Discuss how the utility-maximization model helps highlight the income and substitution effects of a price change.

LO7.5 Give examples of several real-world phenomena that can be explained by applying the theory of consumer behavior.

LO7.6 (Appendix) Relate how the indifference curve model of consumer behavior derives demand curves from budget lines, indifference curves, and utility maximization.

If you were to compare the shopping carts of almost any two consumers, you would observe striking differences. Why does Paula have potatoes, peaches, and Pepsi in her cart, while Sam has sugar, saltines, and 7-Up in his? Why didn't Paula also buy pasta and plums? Why didn't Sam have soup and spaghetti on his grocery list?

In this chapter, you will see how individual consumers allocate their incomes among the various goods and services available to them. Given a certain budget, how does a consumer decide which goods and services to buy? This chapter will develop a model to answer this question.

Law of Diminishing Marginal Utility

LO7.1 Define and explain the relationship between total utility, marginal utility, and the law of diminishing marginal utility.

The simplest theory of consumer behavior rests squarely on the **law of diminishing marginal utility.** This principle, first discussed in Chapter 3, is that added satisfaction declines as a consumer acquires additional units of a given product. Although consumer wants in general may be insatiable, wants for particular items can be satisfied. In a specific span of time over which consumers' tastes remain unchanged, consumers can obtain as much of a particular good or service as they can afford. But the more of that product they obtain, the less they want still more of it.

Consider durable goods, for example. A consumer's desire for an automobile, when he or she has none, may be very strong. But the desire for a second car is less intense; and for a third or fourth, weaker and weaker. Unless they are collectors, even the wealthiest families rarely have more than a half-dozen cars, although their incomes would allow them to purchase a whole fleet of vehicles.

Terminology

Evidence indicates that consumers can fulfill specific wants with succeeding units of a product but that each added unit provides less utility than the last unit purchased. Recall that a consumer derives utility from a product if it can satisfy a want: **Utility** is want-satisfying power. The utility of a good or service is the satisfaction or pleasure one gets from consuming it. Keep in mind three characteristics of this concept:

- "Utility" and "usefulness" are not synonymous. Paintings by Picasso may offer great utility to art connoisseurs but are useless functionally (other than for hiding a crack on a wall).
- Utility is subjective. The utility of a specific product may vary widely from person to person. A lifted pickup truck may have great utility to someone who drives off-road but little utility to someone unable or unwilling to climb into the rig. Eyeglasses have tremendous utility to someone who has poor eyesight but no utility to a person with 20-20 vision.
- Utility is difficult to quantify. But for purposes of illustration we assume that people can measure satisfaction with units called *utils* (units of utility). For example, a particular consumer may get 100 utils of satisfaction from a smoothie, 10 utils of satisfaction from a candy bar, and 1 util of satisfaction from a stick of gum. These imaginary units of satisfaction are convenient for quantifying consumer behavior for explanatory purposes.

Total Utility and Marginal Utility

Total utility and marginal utility are related, but different, ideas. **Total utility** is the total amount of satisfaction or pleasure a person derives from consuming some specific quantity—for example, 10 units—of a good or service. **Marginal utility** is the *extra* satisfaction a consumer realizes from an additional unit of that product—for example, from the eleventh unit. Alternatively, marginal utility is the change in total utility that results from the consumption of 1 more unit of a product.

Figure 7.1(Key Graph) and the accompanying table demonstrate the relation between total utility and marginal utility. The curves reflect the data in the table. Column 2 shows the total utility associated with each level of consumption of tacos. Column 3 shows the marginal utility—the change in total utility—that results from the consumption of each successive taco. Starting at the origin in Figure 7.1a, observe that each of the first five units increases total utility (TU), but by a diminishing amount. Total utility reaches a maximum with the addition of the sixth unit and then declines.

ORIGIN OF THE IDEA

O7.1

Diminishing marginal utility

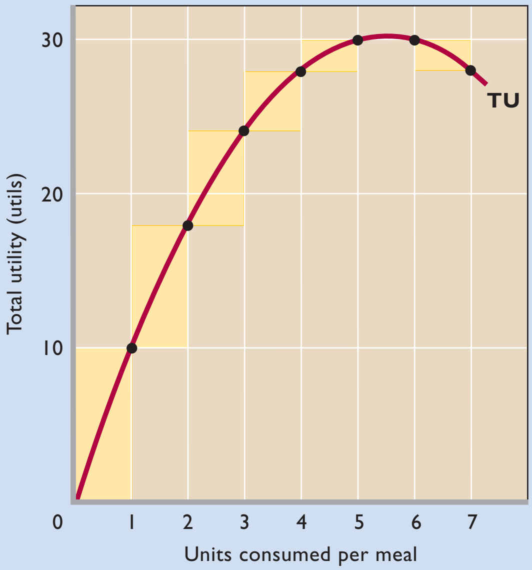
So in Figure 7.1b marginal utility (MU) remains positive but diminishes through the first five units (because total utility increases at a declining rate). Marginal utility is zero for the sixth unit (because that unit doesn't change total utility). Marginal utility then becomes negative with the seventh unit and beyond (because total utility is falling). Figure 7.1b and table column 3 reveal that each successive taco yields less extra utility, meaning fewer utils, than the preceding taco.[1] That is, the table and graph illustrate the law of diminishing marginal utility.

Marginal Utility and Demand

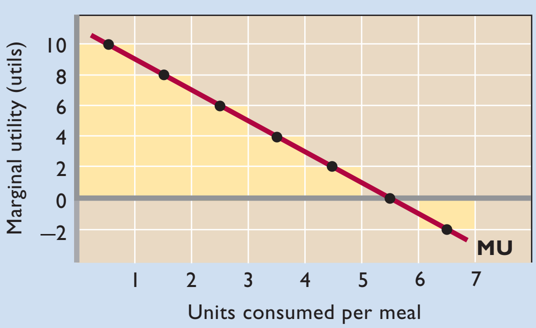
The law of diminishing marginal utility explains why the demand curve for a given product slopes downward. If

[1]Technical footnote: In Figure 7.1b we graphed marginal utility at half-units. For example, we graphed the marginal utility of 4 utils at $3\frac{1}{2}$ units because "4 utils" refers neither to the third nor the fourth unit per se but to the *addition* or *subtraction* of the fourth unit.

KEY GRAPH



(a)
Total utility



(b)
Marginal utility

FIGURE 7.1 Total and marginal utility. Curves TU and MU are graphed from the data in the table. (a) As more of a product is consumed, total utility increases at a diminishing rate, reaches a maximum, and then declines. (b) Marginal utility, by definition, reflects the changes in total utility. Thus marginal utility diminishes with increased consumption, becomes zero when total utility is at a maximum, and is negative when total utility declines. As shown by the shaded rectangles in (a) and (b), marginal utility is the change in total utility associated with each additional taco. Or, alternatively, each new level of total utility is found by adding marginal utility to the preceding level of total utility.

(1) Tacos Consumed per Meal	(2) Total Utility, Utils	(3) Marginal Utility, Utils
0	0	
1	10	10
2	18	8
3	24	6
4	28	4
5	30	2
6	30	0
7	28	–2

QUICK QUIZ FOR FIGURE 7.1

1. Marginal utility:
- **a.** is the extra output a firm obtains when it adds another unit of labor.
- **b.** explains why product supply curves slope upward.
- **c.** typically rises as successive units of a good are consumed.
- **d.** is the extra satisfaction from the consumption of 1 more unit of some good or service.

2. Marginal utility in Figure 7.1b is positive, but declining, when total utility in Figure 7.1a is positive and:
- **a.** rising at an increasing rate.
- **b.** falling at an increasing rate.
- **c.** rising at a decreasing rate.
- **d.** falling at a decreasing rate.

3. When marginal utility is zero in graph (b), total utility in graph (a) is:
- **a.** also zero.
- **b.** neither rising nor falling.
- **c.** negative.
- **d.** rising, but at a declining rate.

4. Suppose the person represented by these graphs experienced a diminished taste for tacos. As a result the:
- **a.** TU curve would get steeper.
- **b.** MU curve would get flatter.
- **c.** TU and MU curves would shift downward.
- **d.** MU curve, but not the TU curve, would collapse to the horizontal axis.

Answers: 1. d; 2. c; 3. b; 4. c

CONSIDER THIS . . .

Vending Machines and Marginal Utility

Newspaper dispensing devices and soft-drink vending machines are similar in their basic operations. Both enable consumers to buy a product by inserting coins. But there is an important difference in the two devices. The newspaper dispenser opens to the full stack of papers and seemingly "trusts" the customer to take only a single copy, whereas the vending machine displays no such "trust," requiring the consumer to buy one can at a time. Why the difference?

The idea of diminishing marginal utility is key to solving this puzzle. Most consumers take only single copies from the newspaper box because the marginal utility of a second newspaper is nearly zero. They could grab a few extra papers and try to sell them on the street, but the revenue obtained would be small relative to their time and effort. So, in selling their product, newspaper publishers rely on "zero marginal utility of the second unit," not on "consumer honesty." Also, newspapers have little "shelf life"; they are obsolete the next day. In contrast, soft-drink sellers do not allow buyers to make a single payment and then take as many cans as they want. If they did, consumers would clean out the machine because the marginal utility of successive cans of soda diminishes slowly and buyers could take extra sodas and consume them later. Soft-drink firms thus vend their products on a pay-per-can basis.

In summary, newspaper publishers and soft-drink firms use alternative vending techniques because of the highly different rates of decline in marginal utility for their products. The newspaper seller uses inexpensive dispensers that open to the full stack of papers. The soft-drink seller uses expensive vending machines that limit the consumer to a single can at a time. Each vending technique is optimal under the particular economic circumstance.

---

successive units of a good yield smaller and smaller amounts of marginal, or extra, utility, then the consumer will buy additional units of a product only if its price falls. The consumer for whom Figure 7.1 is relevant may buy two tacos at a price of $1 each. But because he or she obtains less marginal utility from additional tacos, the consumer will choose not to buy more at that price. The consumer would rather spend additional dollars on products that provide more utility, not less utility. Therefore, additional tacos with less utility are not worth buying unless the price declines. (When marginal utility becomes negative, Taco Bell would have to pay you to consume another taco!) Thus, diminishing marginal utility supports the idea that price must decrease in order for quantity demanded to increase. In other words, consumers behave in ways that make demand curves downsloping.

QUICK REVIEW 7.1

- Utility is the benefit or satisfaction a person receives from consuming a good or a service.
- The law of diminishing marginal utility indicates that gains in satisfaction become smaller as successive units of a specific product are consumed.
- Diminishing marginal utility provides a simple rationale for the law of demand.

Theory of Consumer Behavior

LO7.2 Describe how rational consumers maximize utility by comparing the marginal utility-to-price ratios of all the products they could possibly purchase.

In addition to explaining the law of demand, the idea of diminishing marginal utility explains how consumers allocate their money incomes among the many goods and services available for purchase.

Consumer Choice and the Budget Constraint

For simplicity, we will assume that the situation for the typical consumer has the following dimensions.

- ***Rational behavior*** The consumer is a rational person, who tries to use his or her money income to derive the greatest amount of satisfaction, or utility, from it. Consumers want to get "the most for their money" or, technically, to maximize their total utility. They engage in **rational behavior.**
- ***Preferences*** Each consumer has clear-cut preferences for certain of the goods and services that are available in the market. Buyers also have a good idea of how much marginal utility they will get from successive units of the various products they might purchase.
- ***Budget constraint*** At any point in time the consumer has a fixed, limited amount of money income. Since each consumer supplies a finite amount of human and property resources to society, he or she earns only limited income. Thus, as noted in Chapter 1,

every consumer faces a **budget constraint,** even consumers who earn millions of dollars a year. Of course, this budget limitation is more severe for a consumer with an average income than for a consumer with an extraordinarily high income.

- ***Prices*** Goods are scarce relative to the demand for them, so every good carries a price tag. We assume that the price of each good is unaffected by the amount of it that is bought by any particular person. After all, each person's purchase is a tiny part of total demand. Also, because the consumer has a limited number of dollars, he or she cannot buy everything wanted. This point drives home the reality of scarcity to each consumer.

So the consumer must compromise; he or she must choose the most personally satisfying mix of goods and services. Different individuals will choose different mixes.

Utility-Maximizing Rule

Of all the different combinations of goods and services a consumer can obtain within his or her budget, which specific combination will yield the maximum utility or satisfaction? *To maximize satisfaction, the consumer should allocate his or her money income so that the last dollar spent on each product yields the same amount of extra (marginal) utility.* We call this the **utility-maximizing rule.** When the consumer has "balanced his margins" using this rule, he has achieved **consumer equilibrium** and has no incentive to alter his expenditure pattern. In fact, any person who has achieved consumer equilibrium would be worse off—total utility would decline—if there were any alteration in the bundle of goods purchased, providing there is no change in taste, income, products, or prices.

Numerical Example

An illustration will help explain the utility-maximizing rule. For simplicity we limit our example to two products, but the analysis also applies if there are more. Suppose consumer Holly is analyzing which combination of two products she should purchase with her fixed daily income of $10. Let's suppose these products are apples and oranges.

Holly's preferences for apples and oranges and their prices are the basic data determining the combination that will maximize her satisfaction. Table 7.1 summarizes those data, with column 2a showing the amounts of marginal utility she will derive from each successive unit of A (apples) and with column 3a showing the same thing for product B (oranges). Both columns reflect the law of diminishing marginal utility, which, in this example, is assumed to begin with the second unit of each product purchased.

TABLE 7.1 The Utility-Maximizing Combination of Apples and Oranges Obtainable with an Income of $10*

	(2) Apple (Product A): Price = $1		(3) Orange (Product B): Price = $2	
(1) Unit of Product	(a) Marginal Utility, Utils	(b) Marginal Utility per Dollar (MU/Price)	(a) Marginal Utility, Utils	(b) Marginal Utility per Dollar (MU/Price)
First	10	10	24	12
Second	8	8	20	10
Third	7	7	18	9
Fourth	6	6	16	8
Fifth	5	5	12	6
Sixth	4	4	6	3
Seventh	3	3	4	2

*It is assumed in this table that the amount of marginal utility received from additional units of each of the two products is independent of the quantity of the other product. For example, the marginal-utility schedule for apples is independent of the number of oranges obtained by the consumer.

Marginal Utility per Dollar To see how the utility-maximizing rule works, we must put the marginal-utility information in columns 2a and 3a on a per-dollar-spent basis. A consumer's choices are influenced not only by the extra utility that successive apples will yield but also by how many dollars (and therefore how many oranges) she must give up to obtain additional apples.

The rational consumer must compare the extra utility from each product with its added cost (that is, its price). Switching examples for a moment, suppose that you prefer a pizza whose marginal utility is, say, 36 utils to a movie whose marginal utility is 24 utils. But if the pizza's price is $12 and the movie costs only $6, you would choose the movie rather than the pizza! Why? Because the marginal utility per dollar spent would be 4 utils for the movie (= 24 utils/$6) compared to only 3 utils for the pizza (= 36 utils/$12). You could see two movies for $12 and, assuming that the marginal utility of the second movie is, say, 16 utils, your total utility would be 40 utils. Clearly, 40 units of satisfaction (= 24 utils + 16 utils) from two movies are superior to 36 utils from the same $12 expenditure on one pizza.

To make the amounts of extra utility derived from differently priced goods comparable, marginal utilities must be put on a per-dollar-spent basis. We do this in columns 2b and 3b by dividing the marginal-utility data of columns 2a and 3a by the prices of apples and oranges—$1 and $2, respectively.

TABLE 7.2 Sequence of Purchases to Achieve Consumer Equilibrium, Given the Data in Table 7.1

Choice Number	Potential Choices	Marginal Utility per Dollar	Purchase Decision	Income Remaining
1	First apple	10	First orange for $2	$8 = $10 − $2
	First orange	12		
2	First apple	10	First apple for $1	$5 = $8 − $3
	Second orange	10	and second orange for $2	
3	Second apple	8	Third orange for $2	$3 = $5 − $2
	Third orange	9		
4	Second apple	8	Second apple for $1	$0 = $3 − $3
	Fourth orange	8	and fourth orange for $2	

Decision-Making Process Table 7.1 shows Holly's preferences on a unit basis and a per-dollar basis as well as the price tags of apples and oranges. With $10 to spend, in what order should Holly allocate her dollars on units of apples and oranges to achieve the highest amount of utility within the $10 limit imposed by her income? And what specific combination of the two products will she have obtained at the time she uses up her $10?

Concentrating on columns 2b and 3b in Table 7.1, we find that Holly should first spend $2 on the first orange because its marginal utility per dollar of 12 utils is higher than the first apple's 10 utils. But now Holly finds herself indifferent about whether to buy a second orange or the first apple because the marginal utility per dollar of both is 10 utils per dollar. So she buys both of them. Holly now has 1 apple and 2 oranges. Also, the last dollar she spent on each good yielded the same marginal utility per dollar (10). But this combination of apples and oranges does not represent the maximum amount of utility that Holly can obtain. It cost her only $5 [= (1 × $1) + (2 × $2)], so she has $5 remaining, which she can spend to achieve a still higher level of total utility.

Examining columns 2b and 3b again, we find that Holly should spend the next $2 on a third orange because marginal utility per dollar for the third orange is 9 compared with 8 for the second apple. But now, with 1 apple and 3 oranges, she is again indifferent between a second apple and a fourth orange because both provide 8 utils per dollar. So Holly purchases 1 more of each. Now the last dollar spent on each product provides the same marginal utility per dollar (8), and Holly's money income of $10 is exhausted.

The utility-maximizing combination of goods attainable by Holly is 2 apples and 4 oranges. By summing marginal-utility information from columns 2a and 3a, we find that Holly is obtaining 18 (= 10 + 8) utils of satisfaction from the 2 apples and 78 (= 24 + 20 + 18 + 16) utils of satisfaction from the 4 oranges. Her $10, optimally spent, yields 96 (= 18 + 78) utils of satisfaction.

Table 7.2 summarizes our step-by-step process for maximizing Holly's utility. Note that we have implicitly assumed that Holly spends her entire income. She neither borrows nor saves. However, saving can be regarded as a "commodity" that yields utility and can be incorporated into our analysis. In fact, we treat it that way in problem 4 at the end of this chapter.

Inferior Options Holly can obtain other combinations of apples and oranges with $10, but none will yield as great a total utility as do 2 apples and 4 oranges. As an example, she can obtain 4 apples and 3 oranges for $10. But this combination yields only 93 utils, clearly inferior to the 96 utils provided by 2 apples and 4 oranges. True, there are other combinations of apples and oranges (such as 4 apples and 5 oranges or 1 apple and 2 oranges) in which the marginal utility of the last dollar spent is the same for both goods. But all such combinations either are unobtainable with Holly's limited money income (as 4 apples and 5 oranges) or do not exhaust her money income (as 1 apple and 2 oranges) and therefore do not yield the maximum utility attainable.

WORKED PROBLEMS

W7.1
Consumer choice

Algebraic Generalization

Economists generalize the utility-maximizing rule by saying that a consumer will maximize her satisfaction when she allocates her money income so that the last dollar spent on product A, the last on product B, and so forth, yield equal amounts of additional, or marginal, utility. The marginal utility per dollar spent on A is indicated by the MU of product A divided by the price of A (column 2b in Table 7.1), and the marginal utility per dollar spent on B by the MU of product B divided by the price of B (column 3b in Table 7.1). Our utility-maximizing rule merely requires that these ratios be equal for the last dollar spent on A and the last dollar spent on B. Algebraically,

$$\frac{\text{MU of product A}}{\text{Price of A}} = \frac{\text{MU of product B}}{\text{Price of B}}$$

And, of course, the consumer must exhaust her available income. Table 7.1 shows us that the combination of 2 units of A (apples) and 4 of B (oranges) fulfills these conditions in that

$$\frac{8 \text{ utils}}{\$1} = \frac{16 \text{ utils}}{\$2}$$

and the consumer's $10 income is all spent.

If the equation is not fulfilled, then some reallocation of the consumer's expenditures between A and B (from the low to the high marginal-utility-per-dollar product) will increase the consumer's total utility. For example, if the consumer spent $10 on 4 of A (apples) and 3 of B (oranges), we would find that

$$\frac{\text{MU of A of 6 utils}}{\text{Price of A of \$1}} < \frac{\text{MU of B of 18 utils}}{\text{Price of B of \$2}}$$

Here the last dollar spent on A provides only 6 utils of satisfaction, while the last dollar spent on B provides 9 (= 18/$2). So the consumer can increase total satisfaction by purchasing more of B and less of A. As dollars are reallocated from A to B, the marginal utility per dollar of A will increase while the marginal utility per dollar of B will decrease. At some new combination of A and B the two will be equal and consumer equilibrium will be achieved. Here that combination is 2 of A (apples) and 4 of B (oranges).

Utility Maximization and the Demand Curve

LO7.3 Explain how a demand curve can be derived by observing the outcomes of price changes in the utility-maximization model.

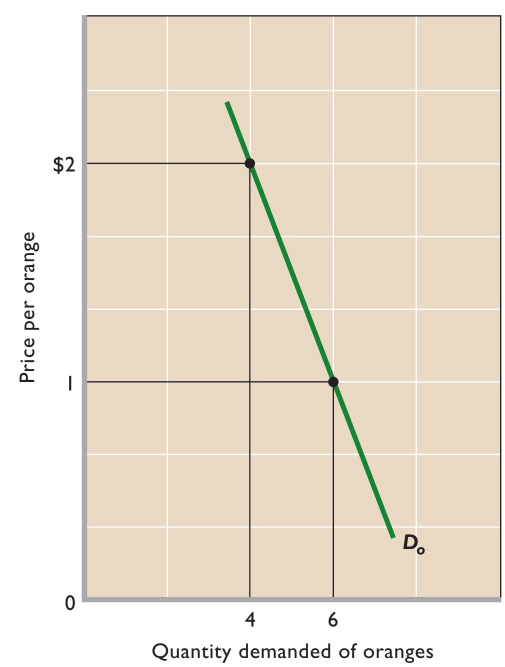
Once you understand the utility-maximizing rule, you can easily see why product price and quantity demanded are inversely related. Recall that the basic determinants of an individual's demand for a specific product are (1) preferences or tastes, (2) money income, and (3) the prices of other goods. The utility data in Table 7.1 reflect our consumer's preferences. We continue to suppose that her money income is $10. And, concentrating on the construction of an individual demand curve for oranges, we assume that the price of apples, now representing all "other goods," is still $1.

Deriving the Demand Schedule and Curve

We can derive a single consumer's demand schedule for oranges by considering alternative prices at which oranges might be sold and then determining the quantity the consumer will purchase. We already know one such price-quantity combination in the utility-maximizing example: Given tastes, income, and the prices of other goods, Holly will purchase 4 oranges at $2.

Now let's assume the price of oranges falls to $1. The marginal-utility-per-dollar data of column 3b in Table 7.1 will double because the price of oranges has been halved; the new data for column 3b are (by coincidence) identical to the data in column 3a. The doubling of the MU per dollar for each successive orange means that the purchase of 2 apples and 4 oranges is no longer an equilibrium combination. By applying the same reasoning we used previously, we now find that Holly's utility-maximizing combination is 4 apples and 6 oranges. As summarized in the table in Figure 7.2,

FIGURE 7.2 Deriving an individual demand curve. The consumer represented by the data in the table maximizes utility by purchasing 4 oranges at a price of $2. The decline in the price of oranges to $1 disrupts the consumer's initial utility-maximizing equilibrium. The consumer restores equilibrium by purchasing 6 rather than 4 oranges. Thus, a simple price-quantity schedule emerges, which locates two points on a downsloping demand curve.



Price per Orange	Quantity Demanded
$2	4
1	6

Holly will purchase 6 oranges when the price of oranges is \$1. Using the data in this table, we can sketch the downward-sloping demand curve for oranges, D_o, shown in Figure 7.2. This exercise, then, clearly links the utility-maximizing behavior of a consumer and that person's downsloping demand curve for a particular product.

Income and Substitution Effects

LO7.4 Discuss how the utility-maximization model helps highlight the income and substitution effects of a price change.

Recall from Chapter 3 that the **income effect** is the impact that a change in the price of a product has on a consumer's real income and consequently on the quantity demanded of that good. In contrast, the **substitution effect** is the impact that a change in a product's price has on its relative expensiveness and consequently on the quantity demanded. Both effects help explain why a demand curve such as that in Figure 7.2 is downsloping.

Let's first look at the substitution effect. Recall that before the price of oranges declined, Holly was in equilibrium when purchasing 2 apples and 4 oranges because

$$\frac{\text{MU of apples of 8}}{\text{Price of apples of \$1}} = \frac{\text{MU of oranges of 16}}{\text{Price of oranges of \$2}}$$

But after the price of oranges declines from \$2 to \$1,

$$\frac{\text{MU of apples of 8}}{\text{Price of apples of \$1}} < \frac{\text{MU of oranges of 16}}{\text{Price of oranges of \$1}}$$

Clearly, the last dollar spent on oranges now yields greater utility (16 utils) than does the last dollar spent on apples (8 utils). This will lead Holly to switch, or substitute, purchases away from apples and toward oranges so as to restore consumer equilibrium. This substitution effect contributes to the inverse relationship between price and quantity that is found along her demand curve for oranges: When the price of oranges declines, the substitution effect causes Holly to buy more oranges.

What about the income effect? The decline in the price of oranges from \$2 to \$1 increases Holly's real income. Before the price decline, she maximized her utility and achieved consumer equilibrium by selecting 2 apples and 4 oranges. But at the lower \$1 price for oranges, Holly would have to spend only \$6 rather than \$10 to buy that particular combination of goods. That means that the lower price of oranges has freed up \$4 that can be spent on buying more apples, more oranges, or more of both. How many more of each fruit she ends up buying will be determined by applying the utility-maximizing rule to the new situation. But it is quite likely that the increase in real income caused by the reduction in the price of oranges will cause Holly to end up buying more oranges than before the price reduction. Any such increase in orange purchases is referred to as the income effect of the reduction in the price of oranges and it, too, helps to explain why demand curves are downward sloping: When the price of oranges falls, the income effect causes Holly to buy more oranges.

ORIGIN OF THE IDEA

O7.2

Income and substitution effects

QUICK REVIEW 7.2

- The theory of consumer behavior assumes that, with limited income and a set of product prices, consumers make rational choices on the basis of well-defined preferences.
- A consumer maximizes utility by allocating income so that the marginal utility per dollar spent is the same for every good purchased.
- A downsloping demand curve can be derived by changing the price of one product in the consumer-behavior model and noting the change in the utility-maximizing quantity of that product demanded.
- By providing insights on the income effect and substitution effect of a price decline, the utility-maximization model helps explain why demand curves are downsloping.

Applications and Extensions

LO7.5 Give examples of several real-world phenomena that can be explained by applying the theory of consumer behavior.

Many real-world phenomena can be explained by applying the theory of consumer behavior.

iPads

Every so often a new product totally captures consumers' imaginations. One such product is Apple's iPad, which debuted in April 2010. Less than three years later, Apple sold its 100-millionth unit.

The swift ascendancy of the iPad resulted mainly from a leapfrog in technology. It was the first touchscreen tablet computer and became a hit because it was much better for the consumption of digital media—music, pictures, videos, and many games—than existing laptop or desktop computers. Those larger machines still held the advantage if a consumer wanted to create content or edit documents, but for consuming digital content the iPad was far superior in the eyes of millions of consumers.

In the language of our analysis, Apple's introduction of the iPad severely disrupted consumer equilibrium. Consumers en masse concluded that iPads had a higher marginal-utility-to-price ratio (= MU/P) than the ratios for alternative products. They therefore shifted spending away from those other products and toward iPads as a way to increase total utility. Of course, for most people the marginal utility of a second or third iPad relative to price is quite low, so most consumers purchased only a single iPad. But Apple continued to enhance the iPad, enticing some of the buyers of older models to buy new models.

This example demonstrates a simple but important point: New products succeed by enhancing consumers' total utility. This "delivery of value" generates a revenue stream. If revenues exceed production costs, substantial profits can result—as they have for Apple.

The Diamond-Water Paradox

Early economists such as Adam Smith were puzzled by the fact that some "essential" goods had much lower prices than some "unimportant" goods. Why would water, essential to life, be priced below diamonds, which have much less usefulness? The paradox is resolved when we acknowledge that water is in great supply relative to demand and thus has a very low price per gallon. Diamonds, in contrast, are rare. Their supply is small relative to demand and, as a result, they have a very high price per carat.

Moreover, the marginal utility of the last unit of water consumed is very low. The reason follows from our utility-maximizing rule. Consumers (and producers) respond to the very low price of water by using a great deal of it—for generating electricity, irrigating crops, heating buildings, watering lawns, quenching thirst, and so on. Consumption is expanded until marginal utility, which declines as more water is consumed, equals its low price. On the other hand, relatively few diamonds are purchased because of their prohibitively high price, meaning that their marginal utility remains high. In equilibrium:

$$\frac{\text{MU of water (low)}}{\text{Price of water (low)}} = \frac{\text{MU of diamonds (high)}}{\text{Price of diamonds (high)}}$$

Although the marginal utility of the last unit of water consumed is low and the marginal utility of the last diamond purchased is high, the total utility of water is very high and the total utility of diamonds quite low. The total utility derived from the consumption of water is large because of the enormous amounts of water consumed. Total utility is the sum of the marginal utilities of all the gallons of water consumed, including the trillions of gallons that have far higher marginal utilities than the last unit consumed. In contrast, the total utility derived from diamonds is low since their high price means that relatively few of them are bought. Thus the water-diamond "paradox" is solved: Water has much more total utility (roughly, usefulness) than diamonds even though the price of diamonds greatly exceeds the price of water. These relative prices relate to marginal utility, not total utility.

ORIGIN OF THE IDEA

O7.3
Diamond-water paradox

Opportunity Cost and the Value of Time

The theory of consumer behavior has been generalized to account for the economic value of *time*. Both consumption and production take time. Time is a valuable economic commodity; by using an hour in productive work a person can earn \$6, \$10, \$50, or more, depending on her or his education and skills. By using that hour for leisure or in consumption activities, the individual incurs the opportunity cost of forgone income; she or he sacrifices the \$6, \$10, or \$50 that could have been earned by working.

Imagine a self-employed consumer named Linden who is considering buying a round of golf, on the one hand, and a concert, on the other. The market price of the golf game is \$30 and that of the concert is \$40. But the golf game takes more time than the concert. Suppose Linden spends 4 hours on the golf course but only 2 hours at the concert. If her time is worth \$10 per hour, as evidenced by the \$10 wage she can obtain by working, then the "full price" of the golf game is \$70 (the \$30 market price plus \$40 worth of time). Similarly, the full price of the concert is \$60 (the \$40 market price plus \$20 worth of time). We find that, contrary to what market prices alone indicate, the full price of the concert is really less than the full price of the golf game.

If we now assume that the marginal utilities derived from successive golf games and concerts are identical, traditional theory would indicate that Linden should consume more golf games than concerts because the market price of the former (\$30) is lower than that of the latter (\$40). But when time is taken into account, the situation is reversed and golf games (\$70) are more expensive than concerts (\$60). So it is rational for Linden to consume more concerts than golf games.

By accounting for the opportunity cost of a consumer's time, we can explain certain phenomena that are otherwise quite puzzling. It may be rational for the unskilled worker

LAST WORD

Criminal Behavior

Although Economic Analysis Is Not Particularly Relevant in Explaining Some Crimes of Passion and Violence (for Example, Murder and Rape), It Does Provide Interesting Insights on Such Property Crimes as Robbery, Burglary, and Auto Theft.

The theory of rational consumer behavior can be extended to provide some useful insights on criminal behavior. Both the lawful consumer and the criminal try to maximize their total utility (or net benefit). For example, you can remove a textbook from the campus bookstore by either purchasing it or stealing it. If you *buy* the book, your action is legal; you have fully compensated the bookstore for the product. (The bookstore would rather have your money than the book.) If you *steal* the book, you have broken the law. Theft is outlawed because it imposes uncompensated costs on others. In this case, your action reduces the bookstore's revenue and profit and also may impose costs on other buyers who now must pay higher prices for their textbooks.

Why might someone engage in a criminal activity such as stealing? Just like the consumer who compares the marginal utility of a good with its price, the potential criminal compares the marginal benefit from his or her action with the "price" or cost. If the marginal benefit (to the criminal) exceeds the price or marginal cost (also to the criminal), the individual undertakes the criminal activity.

Most people, however, do not engage in theft, burglary, or fraud. Why not? The answer is that they perceive the personal price of engaging in these illegal activities to be too high relative to the marginal benefit. The price or marginal cost to the potential criminal has several facets. First, there are the "guilt costs," which for many people are substantial. Such individuals would not steal from others even if there were no penalties for doing so. Their moral sense of right and wrong would entail too great a guilt cost relative to the benefit from the stolen good. Other types of costs include the direct cost of the criminal activity (supplies and tools) and the forgone income from legitimate activities (the opportunity cost to the criminal).

Unfortunately, guilt costs, direct costs, and forgone income are not sufficient to deter some people from stealing. So society imposes other costs, mainly fines and imprisonment, on lawbreakers. The potential of being fined increases the marginal cost to the criminal. The potential of being imprisoned boosts marginal cost still further. Most people highly value their personal freedom and lose considerable legitimate earnings while incarcerated.

Given these types of costs, the potential criminal estimates the marginal cost and benefit of committing the crime. As a simple example, suppose that the direct cost and the opportunity cost of stealing an $80 textbook are both zero. The probability of

getting caught is 10 percent and, if apprehended, there will be a $500 fine. The potential criminal will estimate the marginal cost of stealing the book as $50 (= $500 fine × .10 chance of apprehension). Someone who has a guilt cost of zero will choose to steal the book because the marginal benefit of $80 would exceed the marginal cost of $50. In contrast, someone having a guilt cost of, say, $40 will not steal the book. The marginal benefit of $80 will not be as great as the marginal cost of $90 (= $50 of penalty cost + $40 of guilt cost).

This perspective on illegal behavior has some interesting implications. For example, other things equal, crime will rise (more of it will be "bought") when its price falls. This explains, for instance, why some people who do not steal from stores under normal circumstances participate in looting stores during riots, when the marginal cost of being apprehended has substantially declined.

Another implication is that society can reduce unlawful behavior by increasing the "price of crime." It can nourish and increase guilt costs through family, educational, and religious efforts. It can increase the direct cost of crime by using more sophisticated security systems (locks, alarms, video surveillance) so that criminals will have to buy and use more sophisticated tools. It can undertake education and training initiatives to enhance the legitimate earnings of people who might otherwise engage in illegal activity. It can increase policing to raise the probability of being apprehended for crime. And it can impose greater penalties for those who are caught and convicted.

or retiree whose time has little market value to ride a bus from Chicago to Pittsburgh. But the corporate executive, whose time is very valuable, will find it cheaper to fly, even though bus fare is only a fraction of plane fare. It is sensible for the retiree, living on a modest company pension and a Social Security check, to spend many hours shopping for bargains at the mall or taking long trips in a motor home. It is equally intelligent for the highly paid physician, working 55 hours per week, to buy a new personal computer over the Internet and take short vacations at expensive resorts.

People in other nations often feel affluent Americans are "wasteful" of food and other material goods but "overly economical" in their use of time. Americans who visit developing countries find that time is used casually or "squandered," while material goods are very highly prized and carefully used. These differences are not a paradox or a case of radically different temperaments. The differences are primarily a rational reflection of the fact that the high productivity of labor in an industrially advanced society gives time a high market value, whereas the opposite is true in a low-income, developing country.

Medical Care Purchases

The method of payment for certain goods and services affects their prices at the time we buy them and significantly changes the amount purchased. Let's go back to Table 7.1. Suppose the $1 price for apples is its "true" value or opportunity cost. But now, for some reason, its price is only, say, $0.20. A rational consumer clearly would buy more apples at the $0.20 price than at the $1 price.

That is what happens with medical care. People in the United States who have health insurance pay a fixed premium once a month that covers, say, 80 percent of all incurred health care costs. This means that when they actually need health care, its price to them will be only 20 percent of the actual market price. How would you act in such a situation? When you are ill, you would likely purchase a great deal more medical care than you would if you were confronted with the full price. As a result, financing health care through insurance is an important factor in explaining today's high expenditures on health care and the historical growth of such spending as a percentage of domestic output.

Similar reasoning applies to purchases of buffet meals. If you buy a meal at an all-you-can-eat buffet, you will tend to eat more than if you purchased it item by item. Why not eat that second dessert? Its marginal utility is positive and its "price" is zero!

Cash and Noncash Gifts

Marginal-utility analysis also helps us understand why people generally prefer cash gifts to noncash gifts costing the same amount. The reason is simply that the noncash gifts may not match the recipient's preferences and thus may not add as much as cash to total utility. Thought of differently, consumers know their own preferences better than the gift giver does, and the $100 cash gift provides more choices.

Look back at Table 7.1. Suppose Holly has zero earned income but is given the choice of a $2 cash gift or a noncash gift of 2 apples. Because 2 apples can be bought with $2, these two gifts are of equal monetary value. But by spending the $2 cash gift on the first orange, Holly could obtain 24 utils. The noncash gift of the first 2 apples would yield only 18 (= 10 + 8) units of utility. Conclusion: The noncash gift yields less utility to the beneficiary than does the cash gift.

Since giving noncash gifts is common, a considerable value of those gifts is potentially lost because they do not match their recipients' tastes. For example, Uncle Fred may have paid $15 for the Frank Sinatra CD he gave you for the holidays, but you would pay only $7.50 for it. Thus, a $7.50, or 50 percent, value loss is involved. Multiplied by billions of gifts a year, the total potential loss of value is huge.

But some of that loss is avoided by the creative ways individuals handle the problem. For example, newlyweds set up gift registries for their weddings to help match up their wants to the noncash gifts received. Also, people obtain cash refunds or exchanges for gifts so they can buy goods that provide more utility. And people have even been known to "recycle gifts" by giving them to someone else at a later time. All three actions support the proposition that individuals take actions to maximize their total utility.

SUMMARY

LO7.1 Define and explain the relationship between total utility, marginal utility, and the law of diminishing marginal utility.

The law of diminishing marginal utility states that beyond a certain quantity, additional units of a specific good will yield declining amounts of extra satisfaction to a consumer.

LO7.2 Describe how rational consumers maximize utility by comparing the marginal utility-to-price ratios of all the products they could possibly purchase.

The utility-maximization model assumes that the typical consumer is rational and acts on the basis of well-defined preferences. Because income is limited and goods have prices, the consumer cannot purchase all the goods and services he or she might want. The consumer therefore selects the attainable combination of goods that maximizes his or her utility or satisfaction.

A consumer's utility is maximized when income is allocated so that the last dollar spent on each product purchased yields the same amount of extra satisfaction. Algebraically, the utility-maximizing rule is fulfilled when

$$\frac{\text{MU of product A}}{\text{Price of A}} = \frac{\text{MU of product B}}{\text{Price of B}}$$

and the consumer's total income is spent.

LO7.3 Explain how a demand curve can be derived by observing the outcomes of price changes in the utility-maximization model.

The utility-maximizing rule and the demand curve are logically consistent. Because marginal utility declines, a lower price is needed to induce the consumer to buy more of a particular product.

LO7.4 Discuss how the utility-maximization model helps highlight the income and substitution effects of a price change.

The utility-maximization model illuminates the income and substitution effects of a price change. The income effect implies that a decline in the price of a product increases the consumer's real income and enables the consumer to buy more of that product with a fixed money income. The substitution effect implies that a lower price makes a product relatively more attractive and therefore increases the consumer's willingness to substitute it for other products.

LO7.5 Give examples of several real-world phenomena that can be explained by applying the theory of consumer behavior.

The theory of consumer behavior can explain many real world phenomena, including the rapid adoption of popular consumer goods like the iPad that feature disruptive technologies, the overconsumption of products like health care that have artificially low prices, and why people often prefer gifts of cash to receiving particular items or objects of the same monetary value as gifts.

TERMS AND CONCEPTS

law of diminishing marginal utility
utility
total utility
marginal utility
rational behavior
budget constraint
utility-maximizing rule
consumer equilibrium
income effect
substitution effect

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Complete the following table and answer the questions below: LO7.1

Units Consumed	Total Utility	Marginal Utility
0	0	
1	10	10
2	—	8
3	25	—
4	30	—
5	—	3
6	34	—

   a. At which rate is total utility increasing: a constant rate, a decreasing rate, or an increasing rate? How do you know?
   b. "A rational consumer will purchase only 1 unit of the product represented by these data, since that amount maximizes marginal utility." Do you agree? Explain why or why not.
   c. "It is possible that a rational consumer will not purchase any units of the product represented by these data." Do you agree? Explain why or why not.
2. Mrs. Simpson buys loaves of bread and quarts of milk each week at prices of $1 and 80 cents, respectively. At present she is buying these products in amounts such that the marginal utilities from the last units purchased of the two products are 80 and 70 utils, respectively. Is she buying the utility-maximizing combination of bread and milk? If not, how should she reallocate her expenditures between the two goods? LO7.2
3. How can time be incorporated into the theory of consumer behavior? Explain the following comment: "Want to make millions of dollars? Devise a product that saves Americans lots of time." LO7.2
4. Explain: LO7.2
   a. Before economic growth, there were too few goods; after growth, there is too little time.

b. It is irrational for an individual to take the time to be completely rational in economic decision making.
c. Telling your spouse where you would like to go out to eat for your birthday makes sense in terms of utility maximization.

5. In the last decade or so, there has been a dramatic expansion of small retail convenience stores (such as 7-Eleven, Kwik Shop, and Circle K), although their prices are generally much higher than prices in large supermarkets. What explains the success of the convenience stores? **LO7.2**
6. Many apartment-complex owners are installing water meters for each apartment and billing the occupants according to the amount of water they use. This is in contrast to the former procedure of having a central meter for the entire complex and dividing up the collective water expense as part of the rent. Where individual meters have been installed, water usage has declined 10 to 40 percent. Explain that drop, referring to price and marginal utility. **LO7.3**
7. Using the utility-maximization rule as your point of reference, explain the income and substitution effects of an increase in the price of product B, with no change in the price of product A. **LO7.4**
8. **ADVANCED ANALYSIS** A "mathematically fair bet" is one in which the amount won will on average equal the amount bet, for example, when a gambler bets, say, \$100 for a 10 percent chance to win \$1,000 (\$100 = 0.10 × \$1,000). Assuming diminishing marginal utility of dollars, explain why this is *not* a fair bet in terms of utility. Why is it even a less fair bet when the "house" takes a cut of each dollar bet? So is gambling irrational? **LO7.4**
9. **LAST WORD** In what way is criminal behavior similar to consumer behavior? Why do most people obtain goods via legal behavior as opposed to illegal behavior? What are society's main options for reducing illegal behavior?

REVIEW QUESTIONS

1. True or false. The law of diminishing marginal utility predicts the consumption behavior of addicts quite well. **LO7.1**
2. Frank spends \$75 on 10 magazines and 25 newspapers. The magazines cost \$5 each and the newspapers cost \$2.50 each. Suppose that his MU from the final magazine is 10 utils while his MU from the final newspaper is also 10 utils. According to the utility-maximizing rule, Frank should: **LO7.2**
   a. Reallocate spending from magazines to newspapers.
   b. Reallocate spending from newspapers to magazines.
   c. Be satisfied because he is already maximizing his total utility.
   d. None of the above.
3. Demand curves slope downward because, other things held equal, **LO7.3**
   a. An increase in a product's price lowers MU.
   b. A decrease in a product's price lowers MU.
   c. A decrease in a product's price raises MU per dollar and makes consumers wish to purchase more units.
   d. An increase in a product's price raises MU per dollar and makes consumers wish to purchase more units.
4. Jermaine spends his money on cucumbers and lettuce. If the price of cucumbers falls, the MU per dollar of cucumbers will ____________ and Jermaine will ____________ cucumbers for lettuce. **LO7.4**
   a. Fall; substitute
   b. Rise; substitute
   c. Fall; supply
   d. Rise; demand
5. Tammy spends her money on lemonade and iced tea. If the price of lemonade falls, it is as though her income ____________. **LO7.4**
   a. Increases.
   b. Decreases.
   c. Stays the same.

PROBLEMS

1. Mylie's total utility from singing the same song over and over is 50 utils after one repetition, 90 utils after two repetitions, 70 utils after three repetitions, 20 utils after four repetitions, −50 utils after five repetitions, and −200 utils after six repetitions. Write down her marginal utility for each repetition. Once Mylie's total utility begins to decrease, does each additional singing of the song hurt more than the previous one or less than the previous one? **LO7.1**
2. John likes Coca-Cola. After consuming one Coke, John has a total utility of 10 utils. After two Cokes, he has a total utility of 25 utils. After three Cokes, he has a total utility of 50 utils. Does John show diminishing marginal utility for Coke, or does he show increasing marginal utility for Coke? Suppose that John has \$3 in his pocket. If Cokes cost \$1 each and John is willing to spend one of his dollars on purchasing a first can of Coke, would he spend his second dollar on a Coke, too? What about the third dollar? If John's marginal utility for Coke keeps on increasing no matter how many Cokes he drinks, would it be fair to say that he is addicted to Coke? **LO7.1**
3. Suppose that Omar's marginal utility for cups of coffee is constant at 1.5 utils per cup no matter how many cups he drinks. On the other hand, his marginal utility per doughnut is 10 for the first doughnut he eats, 9 for the second he eats, 8 for the third he eats, and so on (that is, declining by 1 util per additional doughnut). In addition, suppose

Column 1		Column 2		Column 3		Column 4		Column 5	
Units of A	MU	Units of B	MU	Units of C	MU	Units of D	MU	Number of Dollars Saved	MU
1	72	1	24	1	15	1	36	1	5
2	54	2	15	2	12	2	30	2	4
3	45	3	12	3	8	3	24	3	3
4	36	4	9	4	7	4	18	4	2
5	27	5	7	5	5	5	13	5	1
6	18	6	5	6	4	6	7	6	$\frac{1}{2}$
7	15	7	2	7	$3\frac{1}{2}$	7	4	7	$\frac{1}{4}$
8	12	8	1	8	3	8	2	8	$\frac{1}{8}$

that coffee costs \$1 per cup, doughnuts cost \$1 each, and Omar has a budget that he can spend only on doughnuts, coffee, or both. How big would that budget have to be before he would spend a dollar buying a first cup of coffee? **LO7.2**

4. Columns 1 through 4 in the table at the top of the page show the marginal utility, measured in utils, that Ricardo would get by purchasing various amounts of products A, B, C, and D. Column 5 shows the marginal utility Ricardo gets from saving. Assume that the prices of A, B, C, and D are, respectively, \$18, \$6, \$4, and \$24 and that Ricardo has an income of \$106. **LO7.2**

a. What quantities of A, B, C, and D will Ricardo purchase in maximizing his utility?

b. How many dollars will Ricardo choose to save?

c. Check your answers by substituting them into the algebraic statement of the utility-maximizing rule.

5. You are choosing between two goods, X and Y, and your marginal utility from each is as shown in the table to the right. If your income is \$9 and the prices of X and Y are \$2 and \$1, respectively, what quantities of each will you purchase to maximize utility? What total utility will you realize? Assume that, other things remaining unchanged, the price of X falls to \$1. What quantities of X and Y will you now purchase? Using the two prices and quantities for X, derive a demand schedule (a table showing prices and quantities demanded) for X. **LO7.3**

Units of X	MU_x	Units of Y	MU_y
1	10	1	8
2	8	2	7
3	6	3	6
4	4	4	5
5	3	5	4
6	2	6	3

6. **ADVANCED ANALYSIS** Let $MU_A = z = 10 - x$ and $MU_B = z = 21 - 2y$, where z is marginal utility per dollar measured in utils, x is the amount spent on product A, and y is the amount spent on product B. Assume that the consumer has \$10 to spend on A and B—that is, $x + y = 10$. How is the \$10 best allocated between A and B? How much utility will the marginal dollar yield? **LO7.3**

7. Suppose that with a budget of \$100, Deborah spends \$60 on sushi and \$40 on bagels when sushi costs \$2 per piece and bagels cost \$2 per bagel. But then, after the price of bagels falls to \$1 per bagel, she spends \$50 on sushi and \$50 on bagels. How many pieces of sushi and how many bagels did Deborah consume before the price change? At the new prices, how much money would it have cost Deborah to buy those same quantities (the ones that she consumed before the price change)? Given that it used to take Deborah's entire \$100 to buy those quantities, how big is the income effect caused by the reduction in the price of bagels? **LO7.4**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

Indifference Curve Analysis

LO7.6 Relate how the indifference curve model of consumer behavior derives demand curves from budget lines, indifference curves, and utility maximization.

The utility-maximization rule previously discussed requires individuals to measure and compare utility, much as a business would measure and compare costs or revenues. Such *cardinal utility* is measured in units such as 1, 2, 3, and 4 and can be added, subtracted, multiplied, and divided, just like the cardinal numbers in mathematics. More importantly, cardinal utility allows precise quantification of the marginal utilities upon which the utility-maximizing rule depends. In fact, the marginal-utility theory of consumer demand that we explained in the body of this chapter rests squarely on the assumption that economists be able to measure cardinal utility. The reality, however, is that measuring cardinal utility is highly difficult, at best. (Can you, for instance, state exactly how many utils you are getting from reading this book right now or how many utils you would get from watching a sunset?)

To avoid this measurement problem, economists have developed an alternative explanation of consumer behavior and equilibrium in which cardinal measurement is not required. In this more-advanced analysis, the consumer must simply *rank* various combinations of goods in terms of preference. For instance, Sally can simply report that she *prefers* 4 units of A to 6 units of B without having to put number values on how much she likes either option. The model of consumer behavior that is based upon such *ordinal utility* rankings is called indifference curve analysis. It has two main elements: budget lines and indifference curves.

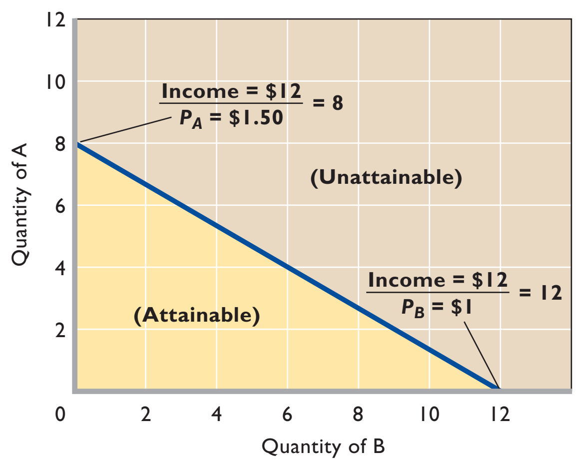
The Budget Line: What Is Attainable

We know from Chapter 1 that a **budget line** (or, more technically, a *budget constraint*) is a schedule or curve showing various combinations of two products a consumer can purchase with a specific money income. If the price of product A is \$1.50 and the price of product B is \$1, a consumer could purchase all the combinations of A and B shown in the table in Figure 1 with \$12 of money income. At one extreme, the consumer might spend all of his or her income on 8 units of A and have nothing left to spend on B. Or, by giving up 2 units of A and thereby "freeing" \$3, the consumer could have 6 units of A and 3 of B. And so on to the other extreme, at which the consumer could buy 12 units of B at \$1 each, spending his or her entire money income on B with nothing left to spend on A.

Figure 1 also shows the budget line graphically. Note that the graph is not restricted to whole units of A and B as is the table. Every point on the graph represents a possible combination of A and B, including fractional quantities. The slope of the graphed budget line measures the ratio of the price of B to the price of A; more precisely, the absolute value of the slope is $P_B/P_A = \$1.00/\$1.50 = \frac{2}{3}$. This is the mathematical way of saying that the consumer must forgo 2 units of

FIGURE 1 A consumer's budget line. The budget line shows all the combinations of any two products that someone can purchase, given the prices of the products and the person's money income.

Units of A (Price = \$1.50)	Units of B (Price = \$1)	Total Expenditure
8	0	\$12 (= \$12 + \$0)
6	3	\$12 (= \$9 + \$3)
4	6	\$12 (= \$6 + \$6)
2	9	\$12 (= \$3 + \$9)
0	12	\$12 (= \$0 + \$12)



A (measured on the vertical axis) to buy 3 units of B (measured on the horizontal axis). In moving down the budget or price line, 2 units of A (at \$1.50 each) must be given up to obtain 3 more units of B (at \$1 each). This yields a slope of $\frac{2}{3}$.

Note that all combinations of A and B that lie on or inside the budget line are attainable from the consumer's \$12 of money income. He can afford to buy not only the combinations of A and B that lie along the budget line itself but also those that lie below it. He could, for instance, afford to buy 2 units of A and 4 units of B, thereby using up only \$7 (= \$3 spent on 2 units of A at a price of \$1.50 each + \$4 spent on 4 units of B at a price of \$1 each). That combination is clearly attainable because it would use up only half of the consumer's \$12 budget. But to achieve maximum utility, the consumer will want to spend the full \$12. The budget line shows all combinations that cost exactly the full \$12.

The budget line has two other significant characteristics:

- ***Income changes*** The location of the budget line varies with money income. An increase in money income shifts the budget line to the right; a decrease in money income shifts it to the left. To verify this, recalculate the table in Figure 1, assuming that money income is (a) \$24 and (b) \$6, and plot the new budget lines in Figure 1.
- ***Price changes*** A change in product prices also shifts the budget line. A decline in the prices of both products—the equivalent of an increase in real income—shifts the curve to the right. (You can verify this by recalculating the table in Figure 1 and replotting Figure 1 assuming that P_A = \$0.75 and P_B= \$0.50.) Conversely, an increase in the prices of A and B shifts the curve to the left. (Assume P_A = \$3 and P_B = \$2, and rework the table and Figure 1 to substantiate this statement.)

Note what happens if P_B changes while P_A and money income remain constant. In particular, if P_B drops, say, from \$1 to \$0.50, the lower end of the budget line fans outward to the right. Conversely, if P_B increases, say, from \$1 to \$1.50, the lower end of the line fans inward to the left. In both instances the line remains "anchored" at 8 units on the vertical axis because P_A has not changed.

Indifference Curves: What Is Preferred

Budget lines reflect "objective" market data, specifically income and prices. They reveal combinations of products A and B that can be purchased, given current money income and prices.

Indifference curves, on the other hand, reflect "subjective" information about consumer preferences for A and B. An **indifference curve** shows all the combinations of two products A and B that will yield the same total satisfaction or total utility to a consumer. The table and graph in Figure 2 present a hypothetical indifference curve for products A and B. The consumer's subjective preferences are such that he or she will realize the same total utility from each combination of A and B shown in the table or on the curve. So the consumer will be indifferent (will not care) as to which combination is actually obtained.

ORIGIN OF THE IDEA

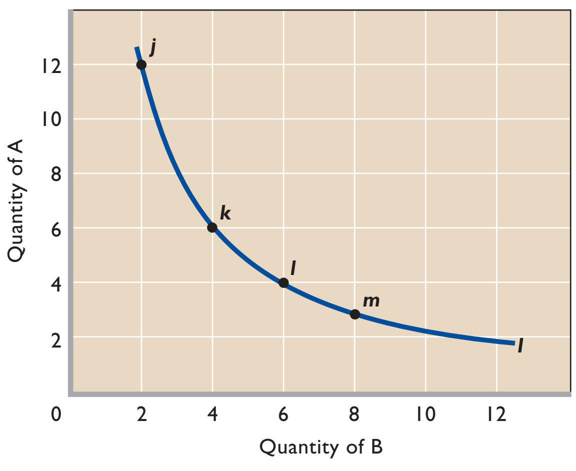
O7.4

Indifference curves

Indifference curves have several important characteristics.

FIGURE 2 A consumer's indifference curve. Every point on indifference curve *I* represents some combination of products A and B, and all those combinations are equally satisfactory to the consumer. That is, each combination of A and B on the curve yields the same total utility.

Combination	Units of A	Units of B
j	12	2
k	6	4
l	4	6
m	3	8



Indifference Curves Are Downsloping

An indifference curve slopes downward because more of one product means less of the other if total utility is to remain unchanged. Suppose the consumer moves from one combination of A and B to another, say, from *j* to *k* in Figure 2. In so doing, the consumer obtains more of product B, increasing his or her total utility. But because total utility is the same everywhere on the curve, the consumer must give up some of the other product, A, to reduce total utility by a precisely offsetting amount. Thus "more of B" necessitates "less of A," and the quantities of A and B are inversely related. A curve that reflects inversely related variables is downsloping.

Indifference Curves Are Convex to the Origin

Recall from the appendix to Chapter 1 that the slope of a curve at a particular point is measured by drawing a straight line that is tangent to that point and then measuring the "rise over run" of the straight line. If you drew such straight lines for several points on the curve in Figure 2, you would find that their slopes decline (in absolute terms) as you move down the curve. An indifference curve is therefore convex (bowed inward) to the origin of the graph. Its slope diminishes or becomes flatter as we move down the curve from *j* to *k* to *l*, and so on. Technically, the slope of an indifference curve at each point measures the **marginal rate of substitution (MRS)** of the combination of two goods represented by that point. The slope or MRS shows the rate at which the consumer who possesses the combination must substitute one good for the other (say, B for A) to remain equally satisfied. The diminishing slope of the indifference curve means that the willingness to substitute B for A diminishes as more of B is obtained.

The rationale for this convexity—that is, for a diminishing MRS—is that a consumer's subjective willingness to substitute B for A (or A for B) will depend on the amounts of B and A he or she has to begin with. Consider the table and graph in Figure 2 again, beginning at point *j*. Here, in relative terms, the consumer has a substantial amount of A and very little of B. Within this combination, a unit of B is very valuable (that is, its marginal utility is high), while a unit of A is less valuable (its marginal utility is low). The consumer will then be willing to give up a substantial amount of A to get, say, 2 more units of B. In this case, the consumer is willing to forgo 6 units of A to get 2 more units of B; the MRS is $\frac{6}{2}$, or 3, for the *jk* segment of the curve.

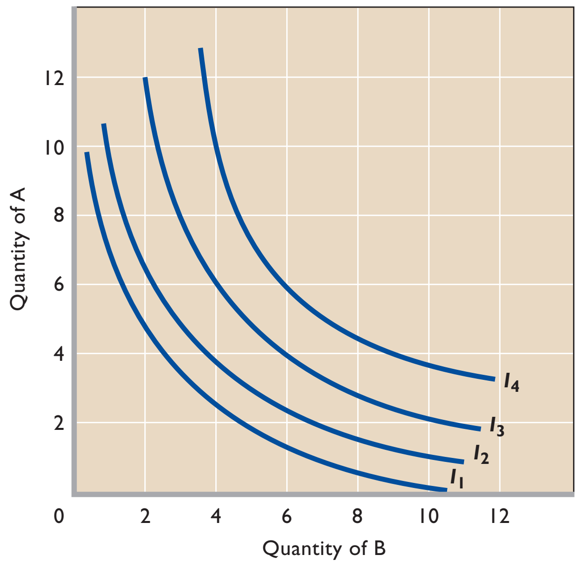
But at point *k* the consumer has less A and more B. Here A is somewhat more valuable, and B less valuable, "at the margin." In a move from point *k* to point *l*, the consumer is willing to give up only 2 units of A to get 2 more units of B, so the MRS is only $\frac{2}{2}$, or 1. Having still less of A and more of B at point *l*, the consumer is willing to give up only 1 unit of A in return for 2 more units of B and the MRS falls to $\frac{1}{2}$ between *l* and *m*.[1]

In general, as the amount of B *increases*, the marginal utility of additional units of B *decreases*. Similarly, as the quantity of A *decreases*, its marginal utility *increases*. In Figure 2 we see that in moving down the curve, the consumer will be willing to give up smaller and smaller amounts of A to offset acquiring each additional unit of B. The result is a curve with a diminishing slope, a curve that is convex to the origin. The MRS declines as one moves southeast along the indifference curve.

The Indifference Map

The single indifference curve of Figure 2 reflects some constant (but unspecified) level of total utility or satisfaction. It is possible and useful to sketch a whole series of indifference curves or an **indifference map,** as shown in Figure 3. Each curve reflects a different level of total utility and therefore never crosses another indifference curve. Specifically, each curve to the right of our original curve (labeled I_3 in

FIGURE 3 An indifference map. An indifference map is a set of indifference curves. Curves farther from the origin indicate higher levels of total utility. Thus any combination of products A and B represented by a point on I_4 has greater total utility than any combination of A and B represented by a point on I_3, I_2, or I_1.



[1]MRS declines continuously between *j* and *k*, *k* and *l*, and *l* and *m*. Our numerical values for MRS relate to the curve segments between points and are not the actual values of the MRS at each point. For example, the MRS *at* point *l* is $\frac{2}{3}$.

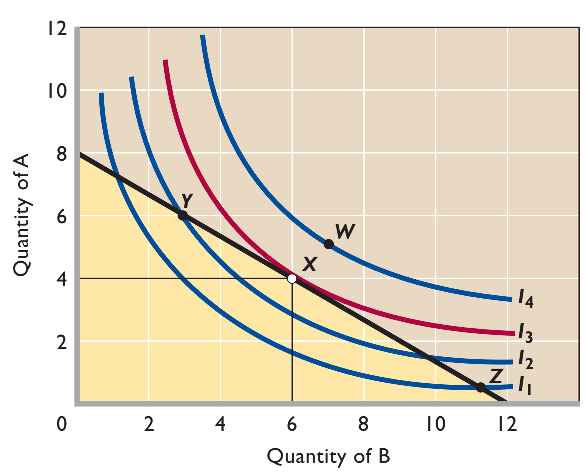
Figure 3) reflects combinations of A and B that yield more utility than I_3. Each curve to the left of I_3 reflects less total utility than I_3. As we move out from the origin, each successive indifference curve represents a higher level of utility. To demonstrate this fact, draw a line in a northeasterly direction from the origin; note that its points of intersection with successive curves entail larger amounts of both A and B and therefore higher levels of total utility.

Equilibrium at Tangency

Since the axes in Figures 1 and 3 are identical, we can superimpose a budget line on the consumer's indifference map, as shown in Figure 4. By definition, the budget line indicates all the combinations of A and B that the consumer can attain with his or her money income, given the prices of A and B. Of these attainable combinations, the consumer will prefer the combination that yields the greatest satisfaction or utility. Specifically, the utility-maximizing combination will be the combination lying on the highest attainable indifference curve. It is called the consumer's **equilibrium position.**

In Figure 4 the consumer's equilibrium position is at point *X*, where the budget line is *tangent* to I_3. Why not point *Y*? Because *Y* is on a lower indifference curve, I_2. By moving "down" the budget line—by shifting dollars from purchases of A to purchases of B—the consumer can attain an indifference curve farther from the origin and thereby increase the total utility derived from the same income. Why not point *Z*? For the same reason: Point *Z* is on a lower indifference curve, I_1. By moving "up" the budget line—by reallocating dollars from B to A—the consumer can get on higher indifference curve I_3 and increase total utility.

FIGURE 4 **The consumer's equilibrium position.** The consumer's equilibrium position is represented by point *X*, where the black budget line is tangent to indifference curve I_3. The consumer buys 4 units of A at \$1.50 per unit and 6 of B at \$1 per unit with a \$12 money income. Points *Z* and *Y* represent attainable combinations of A and B but yield less total utility, as is evidenced by the fact that they are on lower indifference curves. Point *W* would entail more utility than *X*, but it requires a greater income than the \$12 represented by the budget line.



How about point *W* on indifference curve I_4? While it is true that *W* would yield a greater total utility than *X*, point *W* is beyond (outside) the budget line and hence is *not* attainable by the consumer. Point *X* represents the optimal *attainable* combination of products A and B. Note that at the equilibrium position, *X*, the definition of tangency implies that the slope of the highest attainable indifference curve equals the slope of the budget line. Because the slope of the indifference curve reflects the MRS (marginal rate of substitution) and the slope of the budget line is P_B/P_A, the consumer's optimal or equilibrium position is the point where

$$\text{MRS} = \frac{P_B}{P_A}$$

(You may benefit by trying Appendix Discussion Question 3 at this time.)

Equivalency at Equilibrium

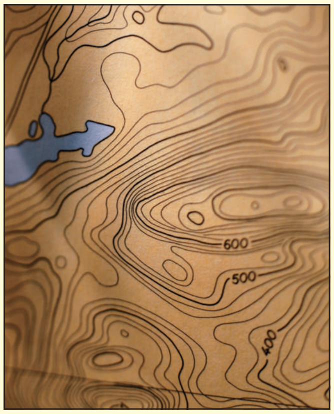
As indicated at the beginning of this appendix, an important difference exists between the marginal-utility theory of consumer demand and the indifference curve theory. The marginal-utility theory assumes that utility is *numerically* measurable, that is, that the consumer can say how much extra utility he or she derives from each extra unit of A or B. The consumer needs that information to determine the utility-maximizing (equilibrium) position, which is defined by

$$\frac{\text{Marginal utility of A}}{\text{Price of A}} = \frac{\text{Marginal utility of B}}{\text{Price of B}}$$

The indifference curve approach imposes a less stringent requirement on the consumer. He or she need only specify whether a particular combination of A and B will yield more than, less than, or the same amount of utility as some other combination of A and B will yield. The consumer need only say, for example, that 6 of A and 7 of B will yield more (or less) satisfaction than will 4 of A and 9 of B. Indifference curve theory does not require that the consumer specify *how much* more (or less) satisfaction will be realized.

That being said, it is a remarkable mathematical fact that both models of consumer behavior will, in any given situation, point to exactly the same consumer equilibrium and, consequently, exactly the same demand behavior. This fact allows us to combine the separate pieces of information that each theory gives us about equilibrium in order to deduce an interesting property about marginal utilities that must also hold true in equilibrium. To see this, note that when we

CONSIDER THIS . . .



Indifference Maps and Topographical Maps

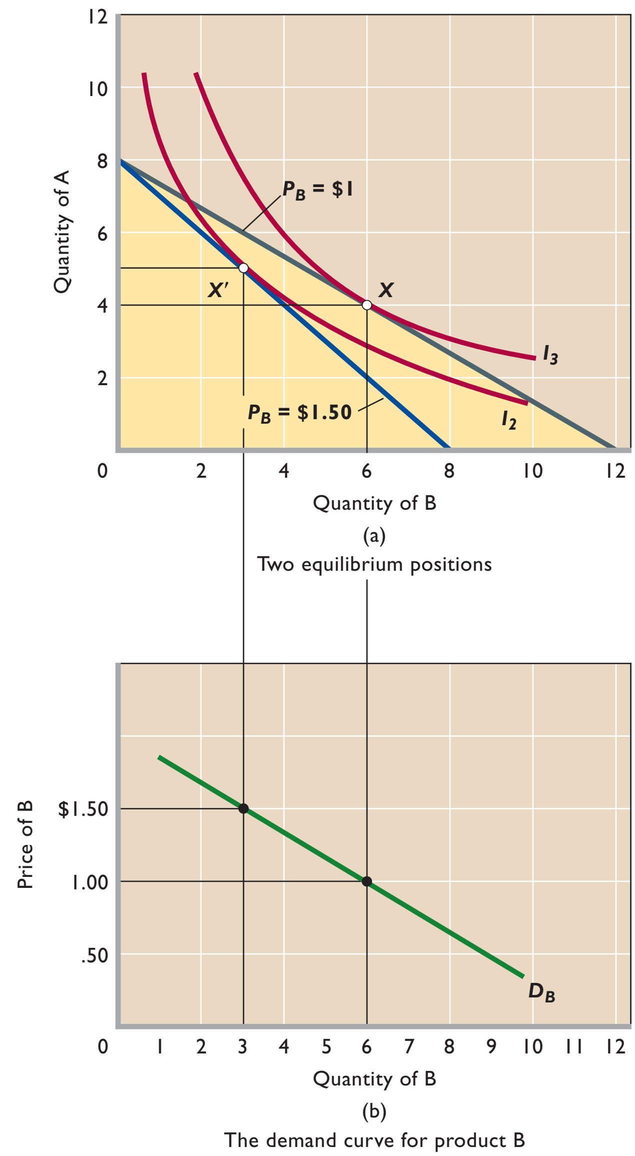
The familiar topographical map may help you understand the idea of indifference curves and indifference maps. Each line on a topographical map represents a particular elevation above sea level, such as 500 feet. Similarly, an indifference curve represents a particular level of total utility. When you move from one point on a specific elevation line to another, the elevation remains the same. So it is with an indifference curve. A move from one position to another on the curve leaves total utility unchanged. Neither elevation lines nor indifference curves can intersect. If they did, the meaning of each line or curve would be violated. An elevation line is "an equal-elevation line"; an indifference curve is "an equal-total-utility curve."

Like the topographical map, an indifference map contains not just one line but a series of lines. That is, the topographical map may have elevation lines representing successively higher elevations of 100, 200, 300, 400, and 500 feet. Similarly, the indifference curves on the indifference map represent successively higher levels of total utility. The climber whose goal is to maximize elevation wants to get to the highest attainable elevation line; the consumer desiring to maximize total utility wants to get to the highest attainable indifference curve.

Finally, both topographical maps and indifference maps show only a few of the many such lines that could be drawn. The topographical map, for example, leaves out the elevation lines for 501 feet, 502, 503, and so on. The indifference map leaves out all the indifference curves that could be drawn between those that are displayed.

compare the equilibrium situations in the two theories, we find that in the indifference curve analysis the MRS equals P_B/P_A at equilibrium; however, in the marginal-utility approach the ratio of marginal utilities equals P_B/P_A. We therefore deduce that at equilibrium the MRS is equivalent in the marginal-utility approach to the ratio of the marginal utilities of the last purchased units of the two products.[2]

FIGURE 5 **Deriving the demand curve.** (a) When the price of product B is increased from \$1 to \$1.50, the equilibrium position moves from *X* to *X′*, decreasing the quantity demanded of product B from 6 to 3 units. (b) The demand curve for product B is determined by plotting the \$1–6-unit and the \$1.50–3-unit price-quantity combinations for product B.



The Derivation of the Demand Curve

We noted earlier that with a fixed price for A, an increase in the price of B will cause the bottom of the budget line to fan inward to the left. We can use that fact to derive a demand curve for product B. In Figure 5a we reproduce the part of Figure 4 that shows our initial consumer equilibrium at point *X*. The budget line determining this

[2]Technical footnote: If we begin with the utility-maximizing rule, $MU_A/P_A = MU_B/P_B$, and then multiply through by P_B and divide through by MU_A, we obtain $P_B/P_A = MU_B/MU_A$. In indifference curve analysis we know that at the equilibrium position MRS $= P_B/P_A$. Hence, at equilibrium, MRS also equals MU_B/MU_A.

equilibrium position assumes that money income is $12 and that P_A = $1.50 and P_B = $1. Let's see what happens to the equilibrium position when we increase P_B to $1.50 and hold both money income and the price of A constant. The result is shown in Figure 5a. The budget line fans to the left, yielding a new equilibrium point X' where it is tangent to lower indifference curve I_2. At X' the consumer buys 3 units of B and 5 of A, compared with 4 of A and 6 of B at X. Our interest is in B, and we now have sufficient information to locate two points on the demand curve for product B. We know that at equilibrium point X the price of B is $1 and 6 units are purchased; at equilibrium point X' the price of B is $1.50 and 3 units are purchased.

These data are shown graphically in Figure 5b as points on the consumer's demand curve for B. Note that the horizontal axes of Figures 5a and 5b are identical; both measure the quantity demanded of B. We can therefore drop vertical reference lines from Figure 5a down to the horizontal axis of Figure 5b. On the vertical axis of Figure 5b we locate the two chosen prices of B. Knowing that these prices yield the relevant quantities demanded, we locate two points on the demand curve for B. By simple manipulation of the price of B in an indifference curve–budget line context, we have obtained a downward-sloping demand curve for B. We have thus again derived the law of demand assuming "other things equal," since only the price of B was changed (the price of A and the consumer's money income and tastes remained constant). But, in this case, we have derived the demand curve without resorting to the questionable assumption that consumers can measure utility in units called "utils." In this indifference curve approach, consumers simply compare combinations of products A and B and determine which combination they prefer, given their incomes and the prices of the two products.

APPENDIX SUMMARY

LO7.6 Relate how the indifference curve model of consumer behavior derives demand curves from budget lines, indifference curves, and utility maximization.

The indifference curve approach to consumer behavior is based on the consumer's budget line and indifference curves.

The budget line shows all combinations of two products that the consumer can purchase, given product prices and his or her money income. A change in either product prices or money income moves the budget line.

An indifference curve shows all combinations of two products that will yield the same total utility to a consumer. Indifference curves are downsloping and convex to the origin.

An indifference map consists of a number of indifference curves; the farther from the origin, the higher the total utility associated with a curve.

The consumer is in equilibrium (utility is maximized) at the point on the budget line that lies on the highest attainable indifference curve. At that point the budget line and indifference curve are tangent.

Changing the price of one product shifts the budget line and determines a new equilibrium point. A downsloping demand curve can be determined by plotting the price-quantity combinations associated with two or more equilibrium points.

APPENDIX TERMS AND CONCEPTS

budget line
indifference curve
marginal rate of substitution (MRS)
indifference map
equilibrium position

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. What information is embodied in a budget line? What shifts occur in the budget line when money income (*a*) increases and (*b*) decreases? What shifts occur in the budget line when the price of the product shown on the vertical axis (*c*) increases and (*d*) decreases? **LO7.6**

2. What information is contained in an indifference curve? Why are such curves (*a*) downsloping and (*b*) convex to the origin? Why does total utility increase as the consumer moves to indifference curves farther from the origin? Why can't indifference curves intersect? **LO7.6**
3. Using Figure 4, explain why the point of tangency of the budget line with an indifference curve is the consumer's equilibrium position. Explain why any point where the budget line intersects an indifference curve is not equilibrium. Explain: "The consumer is in equilibrium where MRS = P_B/P_A." **LO7.6**

APPENDIX REVIEW QUESTIONS

1. Consider two bundles of coffee and chocolate and how Ted feels about them. The first bundle consists of two cups of coffee and two chocolate bars. The second bundle consists of one cup of coffee and three chocolate bars. If the first bundle gives Ted a total utility of 18 utils while the second bundle gives Ted a total utility of 19 bundles, could the two bundles be on the same indifference curve? Answer yes or no. **LO7.6**
2. Bill spends his money on flowers and cookies so as to maximize his total utility. Both flowers and cookies start off costing $2 each. At that price, Bill buys three flowers and two cookies. When the price of flowers is lowered to $1, Bill buys eight flowers and one cookie. Which of the following statements about Bill's reaction to the price change is **not** true? **LO7.6**
   - **a.** Bill's budget line shifted outward when the price of flowers fell.
   - **b.** Bill moved to a higher indifference curve after the price of flowers fell.
   - **c.** Bill's demand curve for flowers shifted to the right.
   - **d.** Bill's attainable set was smaller before the price of flowers fell.

APPENDIX PROBLEMS

1. Assume that the data in the accompanying table give an indifference curve for Mr. Chen. Graph this curve, putting A on the vertical axis and B on the horizontal axis. Assuming that the prices of A and B are $1.50 and $1, respectively, and that Mr. Chen has $24 to spend, add his budget line to your graph. What combination of A and B will Mr. Chen purchase? Does your answer meet the MRS = P_B/P_A rule for equilibrium? **LO7.6**

Units of A	Units of B
16	6
12	8
8	12
4	24

2. Explain graphically how indifference analysis can be used to derive a demand curve. **LO7.6**
3. **ADVANCED ANALYSIS** First, graphically illustrate a doubling of income without price changes in the indifference curve model. Next, on the same graph, show a situation in which the person whose indifference curves you are drawing buys considerably more of good B than good A after the income increase. What can you conclude about the relative coefficients of the income elasticity of demand for goods A and B (Chapter 6)? **LO7.6**

CHAPTER **8**

Behavioral Economics

Learning Objectives

LO8.1 **Define behavioral economics and explain how it contrasts with neoclassical economics.**

LO8.2 **Discuss the evidence for the brain being modular, computationally restricted, reliant on heuristics, and prone to various forms of cognitive error.**

LO8.3 **Relate how prospect theory helps to explain many consumer behaviors, including framing effects, mental accounting, anchoring, loss aversion, and the endowment effect.**

LO8.4 **Describe how time inconsistency and myopia cause people to make suboptimal long-run decisions.**

LO8.5 **Define fairness and give examples of how it affects behavior in the economy and in the dictator and ultimatum games.**

Scientific theories are judged by the accuracy of their predictions. As an example, nobody would take physics seriously if it weren't possible to use the equations taught in college physics classes to predict the best trajectory for putting a satellite into orbit or the best radio frequency to penetrate buildings and provide good indoor cellular service.

Conventional **neoclassical economics** makes many accurate predictions about human choice behavior, especially when it comes to financial incentives and how consumers and businesses respond to changing prices. On the other hand, a number of neoclassical predictions fail quite dramatically. These include predictions about how people deal with risk and uncertainty; choices that require willpower or commitment; and decisions that involve fairness, reciprocity, or trust.

Behavioral economics attempts to make better predictions about human choice behavior by combining insights from economics, psychology, and biology. This chapter introduces you to behavioral economics and the areas in which it has most dramatically increased our understanding of economic behavior. Among the highlights is prospect theory, which was such a large advance on our understanding of how people deal with risk and uncertainty that its inventor, Daniel Kahneman, received the Nobel Prize in economics.

Systematic Errors and the Origin of Behavioral Economics

LO8.1 Define behavioral economics and explain how it contrasts with neoclassical economics.

We tend to think of ourselves as being very good at making decisions. While we may make a few mistakes here and there, we generally proceed through life with confidence, believing firmly that we will react sensibly and make good choices whenever decisions have to be made. In terms of economic terminology, we feel that our decisions are **rational,** meaning that they maximize our chances of achieving what we want.

Unfortunately, scientists have amassed overwhelming evidence to the contrary. People constantly make decision errors that reduce—rather than enhance—the likelihood of getting what they want. In addition, many errors are **systematic errors,** meaning that people tend to repeat them over and over, no matter how many times they encounter a similar situation.

Behavioral economics developed as a separate field of study because neoclassical economics could not explain why people make so many systematic errors. The underlying problem for neoclassical economics is that it assumes that people are fundamentally rational. Under that worldview, people might make some initial mistakes when encountering a new situation. But as they gain experience, they should learn and adapt to the situation. As a result, decision errors should be rare—and definitely not systematic or regularly repeated.

When evidence began to pile up in the late 20th century that even highly experienced people made systematic errors, neoclassical economists assumed that people were just ignorant of what was in their best interests. They assumed that a little education would fix everything. But people often persisted in making the same error even after they were informed that they were behaving against their own interests.

As a result, several researchers realized that it would be necessary to drop the neoclassical assumption that people are fundamentally rational. With that assumption relaxed, economists could develop alternative theories that could make more accurate predictions about human behavior—including the tendency people have toward making systematic errors in certain situations. The result of those efforts is what we today refer to as behavioral economics. Its distinguishing feature is that it is based upon people's actual behavior—which is in many cases substantially irrational, prone to systematic errors, and difficult to modify.

Comparing Behavioral Economics with Neoclassical Economics

While rationality is the most fundamental point of disagreement between behavioral economics and neoclassical economics, it is not the only one. Behavioral economics also contends that neoclassical economics makes a number of highly unrealistic assumptions about human capabilities and motivations, including

- People have stable preferences that aren't affected by context.
- People are eager and accurate calculating machines.
- People are good planners who possess plenty of willpower.
- People are almost entirely selfish and self-interested.

Neoclassical economics made these "simplifying assumptions" for two main reasons. First, they render neoclassical models of human behavior both mathematically elegant and easy enough to solve. Second, they enable neoclassical models to generate very precise predictions about human behavior.

Unfortunately, precision is not the same thing as accuracy. As noted behavioral economist Richard Thaler has written, "Would you rather be elegant and precisely wrong—or messy and vaguely right?"

Behavioral economists err on the side of being messy and vaguely right. As a result, behavioral economics replaces the simplifying assumptions made by neoclassical economics with much more realistic and complex models of human capabilities, motivations, and mental processes.

TABLE 8.1 **Major Differences between Behavioral Economics and Conventional Neoclassical Economics**

Topic	Neoclassical Economics	Behavioral Economics
Rationality	People are fundamentally rational and will adjust their choices and behaviors to best achieve their goals. Consequently, they will not make systematic errors.	People are irrational and make many errors that reduce their chances of achieving their goals. Some errors are regularly repeated systematic errors.
Stability of preferences	People's preferences are completely stable and unaffected by context.	People's preferences are unstable and often inconsistent because they depend on context (framing effects).
Capability for making mental calculations	People are eager and accurate calculators.	People are bad at math and avoid difficult computations if possible.
Ability to assess future options and possibilities	People are just as good at assessing future options as current options.	People place insufficient weight on future events and outcomes.
Strength of willpower	People have no trouble resisting temptation.	People lack sufficient willpower and often fall prey to temptation.
Degree of selfishness	People are almost entirely self-interested and self-centered.	People are often selfless and generous.
Fairness	People do not care about fairness and only treat others well if doing so will get them something they want.	Many people care deeply about fairness and will often give to others even when doing so will yield no personal benefits.

Table 8.1 summarizes how the two approaches differ in several areas.

Focusing on the Mental Processes behind Decisions Another major difference between behavioral economics and neoclassical economics is in the amount of weight and importance that they attach to predicting decisions on the one hand and in understanding the mental processes used to reach those decisions on the other. While neoclassical economics focuses almost entirely on predicting behavior, behavioral economics puts significant emphasis on the mental processes driving behavior.

Neoclassical economics focuses its attention on prediction because its assumption that people are rational allows it to fully separate what people do from how they do it. In particular, perfectly rational people will always choose the course of action that will maximize the likelihood of getting what they want. How they actually come to those optimal decisions might be interesting—but you don't need to know anything about that process to predict a perfectly rational person's behavior. He will simply end up doing whatever it is that will best advance his interests. Consequently, neoclassical economists have felt free to ignore the underlying mental processes by which people make decisions.

Behavioral economists disagree sharply with the neoclassical neglect of mental processes. To them, the fact that people are not perfectly rational implies two important reasons for understanding the underlying mental processes that determine decisions:

- It should allow us to make better predictions about behavior.
- It should provide guidance about how to get people to make better decisions.

Improving Outcomes by Improving Decision-Making Neoclassical economics and behavioral economics differ on how to improve human welfare. Neoclassical economics focuses its attention on providing people with more options. That's because a fully rational person can be trusted to select from any set of options the one that will make him best off. As a result, the only way to make him even happier would be to provide an additional option that is even better.

By contrast, the existence of irrationality leads behavioral economists to conclude that it may be possible to make people better off without providing additional options. In particular, improvements in utility and happiness may be possible simply by getting people to make better selections from the set of options that is already available to them.

This focus on improving outcomes by improving decisions is one of the distinguishing characteristics of behavioral economics. This chapter's Last Word reviews several instances where substantial benefits arise from helping people to make better choices from among the options that they already have.

Viewing Behavioral Economics and Neoclassical Economics as Complements It would be hasty to view behavioral economics and neoclassical economics as fundamentally opposed or mutually exclusive. Instead, many economists prefer to think of them as complementary approaches that can be used in conjunction to help improve our understanding of human behavior.

As an example of their complementary nature, consider how using the two approaches in tandem can help us achieve a better understanding of how customers behave at a local supermarket.

Neoclassical Economics at the Supermarket The major neoclassical contribution to our understanding of the customers' shopping behavior can be summarized by the phrase "incentives matter." In particular, the customers will care a great deal about prices. When prices go up, they buy less. When prices go down, they buy more.

That insight goes a long way toward explaining how customers behave. But there are other shopping behaviors that neoclassical economics cannot explain with its emphasis on people reacting rationally to incentives and prices. In those cases, behavioral economics may be able to help us figure out what people are up to.

Behavioral Economics at the Supermarket A good example of a shopping behavior that neoclassical economics can't explain very well is that people tend to buy what they happen to see. This behavior is called *impulse buying* and it contradicts the neoclassical assumption that consumers carefully calculate marginal utilities and compare prices before making their purchases. On the other hand, it is a very common behavior that is regularly exploited by retailers.

For instance, nearly all supermarkets attempt to take advantage of impulse buying by placing staple products like milk and eggs against the back walls of their stores. Placing those products at the rear increases impulse buying by forcing customers to walk past hundreds of other items on the way to the milk and eggs. A few of those items will catch their eyes and thereby increase sales as customers end up purchasing products that they had no intention of buying when they first entered the store.

Marketers also know that impulse purchases are highest for items that are stacked on shelves at eye level. So, believe it or not, food manufacturers actively bid against each other and pay supermarkets for the privilege of having their brands stacked at eye level. In cereal aisles, the most expensive shelf space isn't at eye level for an adult, but a foot or two lower—at the eye level of a toddler sitting in a shopping cart or of a child walking with a parent. Because kids are even more prone to impulse buying than adults, cereal makers are more than happy to pay to have their products stacked at kid-friendly eye levels.

Complementary Explanations at the Supermarket Behavioral economics explains impulse buying and other irrational behaviors as the result of a wide variety of underlying factors, including cognitive biases, heuristics, and ongoing battles between different areas of the brain.

You will learn about these underlying factors in the remainder of this chapter. But for now, take to heart the idea that we typically need both neoclassical *and* behavioral methods to figure out what people are doing. Some behaviors—including the fact that shoppers respond strongly to incentives and prices—can be explained very well by neoclassical models that assume people are perfectly rational. But other behaviors—including impulse buying—are very much inconsistent with rationality and are therefore better explained by using the methods of behavioral economics.

CONSIDER THIS . . .

Wannamaker's Lament

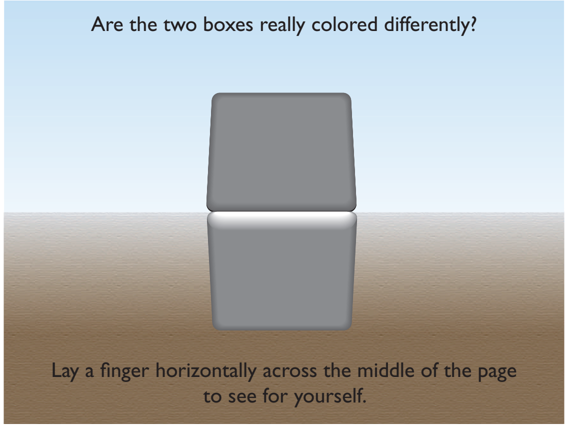
Marketing experts try to increase sales or launch new products by applying what they think they know about consumer behavior. Many people find those efforts spooky and wonder if they are being constantly manipulated into purchasing products that they don't want. But how much do the marketing experts really know?

Judging by their success rate, not so much. Most advertising campaigns show little effect on sales. Eighty percent of newly launched consumer products fail within just three months. And the vast majority of Hollywood films end up as flops despite studios spending billions of dollars each year on market research and advertising.

The difficulties facing marketers were best described in the late 19th century by John Wannamaker, the marketing genius and department store entrepreneur who, among other things, invented the price tag and the money-back guarantee. He famously complained, "Half the money I spend on advertising is wasted—the trouble is, I don't know which half!"

A recent response to Wannamaker's lament has been to run lots of simple experiments to see if anything at all can increase sales. Amazon.com runs hundreds of experiments per month, systematically showing different groups of customers different versions of its website in order to see if any of those different versions can increase sales. Las Vegas casinos also run experiments, systematically varying the scents injected into their air-conditioning systems to see which ones cause the largest increases in gambling. Vanilla apparently works very well and some scents are said to increase revenues by up to 20 percent.

FIGURE 8.1 A visual illusion. The human brain uses a large number of heuristics (shortcuts) to process both visual and other types of information. Many of them utilize context to interpret specific bits of information. When that context changes (as it does here when you put your finger horizontally across the middle of the image), so does the brain's heuristic-filtered interpretation.



Our Efficient, Error-Prone Brains

LO8.2 Discuss the evidence for the brain being modular, computationally restricted, reliant on heuristics, and prone to various forms of cognitive error.

The human brain is the most complex object in the universe. One hundred billion neurons share 10,000 times as many connections. Working together, they allow you to observe your environment, think creatively, and interact with people and objects.

The brain, however, is rather error-prone. Its many weaknesses are most dramatically illustrated by visual illusions, such as the one shown in Figure 8.1. If you follow the instructions printed in that figure, you will quickly discover that your brain can't consistently tell what color an object is.

This inability to properly process visual information is especially informative about the brain's limitations because the brain devotes more neurons toward processing and interpreting visual information than it does anything else. So, if the brain makes errors with visual processing, we should expect to find errors in everything else it does, too.

Heuristics Are Energy Savers

The brain's information-processing limitations are the result of evolutionary pressures. In particular, it was normally very difficult for our ancestors to get enough food to eat. That matters because our brains are extremely energy intensive. In fact, while your brain accounts for just five percent of your body weight, it burns 20 percent of all the energy you consume each day. So back when our ancestors had to hunt and gather and scavenge to survive, getting enough energy was a constant challenge.

In response, the brain evolved many low-energy mental shortcuts, or **heuristics.** Because they are shortcuts, heuristics are not the most accurate mental-processing options. But in a world where calories were hard to come by, a low-energy "good enough" heuristic was superior to a "perfect but costly" alternative.

Your brain's susceptibility to the visual-processing failure demonstrated in Figure 8.1 is the result of your brain using a host of error-prone heuristics to process visual information. But think about what a good trade-off you are getting. In everyday life, the visual-processing failure demonstrated in Figure 8.1 hardly ever comes up. So, it would be a waste of resources to devote more brainpower to fixing the issue. Put in economic terms, there are diminishing returns to employing additional units of brainpower. Heuristics are used because the opportunity cost of perfection is too high.

Some Common Heuristics The following examples will give you a sense of how the brain employs heuristics for nearly every type of action and decision we make.

Catching a Baseball with the Gaze Heuristic Consider the problem faced by a centerfielder in a baseball game when a ball is hit in his general direction. The mentally expensive way to catch the ball would be for the player to use the laws of physics to determine where the ball is heading so that he could run to that spot before the ball arrives.

What baseball players actually do is lock their eyes on the ball and then adjust their position on the field as necessary to keep the ball in front of them and at the same angle as when they first locked their eyes on it. Just as long as they can run fast enough, this *gaze heuristic* always gets them to the correct place to make the catch. You don't need to learn physics to catch a baseball!

Riding a Bicycle with the Steering Heuristic There is a simple heuristic for staying upright as you ride a bicycle: if you begin to fall, steer in the direction you are falling. This *steering heuristic* works because turning in the direction of a fall generates a centrifugal force that can be used to hold you up long enough for you to steady the bike. This heuristic is almost never articulated, but it is precisely what little kids subconsciously learn to do when they are using training wheels.

Guesstimating Ranks with the Recognition Heuristic Which German city has the larger population, Munich or Stuttgart?

Even people who know nothing about Germany tend to get the right answer to this question. They correctly guess "Munich" by employing the *recognition heuristic*, which says to assume that if one option is more easily recognized, it is probably more important or more highly ranked.

The recognition heuristic isn't foolproof, but it tends to work because relatively important people and places are much more likely to be mentioned in the media. Thus, whichever option is easier to recognize will probably be larger or more important.

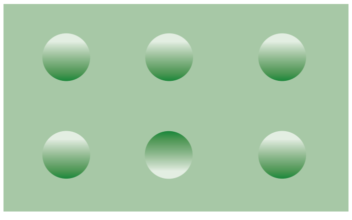
Much of advertising is based on exploiting the recognition heuristic. Indeed, companies spend billions to ensure that consumers are familiar with their products because when it comes time to buy, consumers will be biased toward the products that seem the most familiar.

Interpreting Depth with the Shadow Heuristic The world is three-dimensional, but the light-sensing surfaces at the back of our eyes are two-dimensional. As a result, our brains are forced to use a cluster of heuristics to estimate depth when interpreting the two-dimensional images registered by our eyes.

Figure 8.2 shows how the *shadow heuristic* causes you to interpret shaded, two-dimensional circles as either humps or holes depending upon whether each circle is shaded on the top or on the bottom. Look at Figure 8.2 and count how many of the six shaded circles look like humps rather than holes. Now turn the picture upside down and count again. If your vision is typical, you will find that all the humps have become holes, and vice versa.

Here's what's happening. The shadow heuristic evolved back when sunlight was the only important source of light. As a result, it presumes that light always falls from above. Under that assumption, anything that sticks out from a surface will cast a shadow below it while anything indented will have a shadow on top due to the top of the recessed area casting a shadow on whatever lies below.

FIGURE 8.2 The shadow heuristic. The brain processes light with a heuristic that assumes that light always comes from above. Under that assumption, anything that sticks out will have a shadow on the bottom while anything that is recessed will have a shadow on top. As a result, your brain interprets five of the six shaded circles as humps that stick out while the bottom middle circle is interpreted as a recess. See what happens when you turn the picture upside down. Surprised?



Because your brain applies the shadow heuristic no matter what, you are tricked into believing that the shaded circles in Figure 8.2 are three-dimensional and either humps or holes depending upon whether they are shaded on the top or on the bottom.

The Implications of Hardwired Heuristics As you study the rest of the chapter, keep in mind that most heuristics appear to be hardwired into the brain, and, consequently, impossible to unlearn or avoid. That possibility has three important implications:

1. It may be very difficult for people to alter detrimental behaviors or routines even after you point out what they're doing wrong.
2. People may be easy prey for those who understand their hardwired tendencies.
3. If you want people to make a positive behavioral change, it might be helpful to see if you can put them in a situation where a heuristic will kick in and subconsciously lead them toward the desired outcome.

Brain Modularity

The modern human brain is modular, so that specific areas deal with specific sensations, activities, and emotions—such as vision, breathing, and anger.

This modular structure is the result of millions of years of evolution, with the modern human brain evolving in stages from the much less complex brains of our hominid ancestors. The oldest parts of the brain are located in the back of the head, where the spine enters the skull. The newest parts are up front, near the forehead.

The older parts control subconscious activities like breathing and sweating as well as automatic emotional reactions such as fear and joy. The newer parts allow you to think creatively, imagine the future, and keep track of everyone in your social network. They are largely under conscious control.

System 1 and System 2 It is useful to think of the brain's decision-making systems as falling into two categories, which are informally referred to as System 1 and System 2. System 1 uses a lot of heuristics in the older parts of your brain to produce quick, unconscious reactions. If you ever get a "gut instinct," System 1 is responsible. By contrast, System 2 uses the newer parts of your brain to undertake slow, deliberate, and conscious calculations of costs and benefits. If you ever find yourself "thinking things over," you are using System 2.

Conflicts may sometimes arise between our unconscious System 1 intuitions and our conscious System 2 deliberations. For example, System 1 may urge you to eat an entire pile of cookies as fast as possible, while System 2 admonishes you to stick to your diet and have only one. That being said, a large body of evidence suggests that most decisions are probably either fully or mostly the result of System 1 intuitions and heuristics. That matters because those unconscious mental processes suffer from a variety of cognitive biases.

Cognitive Biases **Cognitive biases** are the misperceptions or misunderstandings that cause systematic errors.

There are a wide variety of cognitive biases, but they can be placed into two general categories. The first are mental-processing errors that result from faulty heuristics. As previously discussed, faulty heuristics are the result of evolution trading off accuracy for speed and efficiency.

The second category of cognitive biases consists of mental-processing errors that result from our brains not having any evolved capacities for dealing with modern problems and challenges, such as solving calculus problems or programming computers. Because our ancestors never encountered things like math, engineering, or statistics, our brains have a total absence of System 1 heuristics for dealing with those sorts of problems. In addition, our slower and more deliberative System 2 mental processes are also of only limited assistance because they were evolved to deal with other types of problems, such as keeping track of everyone in a social network or attempting to think through whether it would be better to go hunting in the morning or in the evening.

As a result, most people find recently developed mental challenges like math and physics to be very tiresome. In addition, cognitive biases often result because the System 2 processes that we are recruiting to solve modern problems were in fact designed for other purposes and don't work particularly well when directed at modern problems.

Psychologists have identified scores of cognitive biases. Here are a few that are relevant to economics and decision-making.

Confirmation Bias The term *confirmation bias* refers to the human tendency to pay attention only to information that agrees with one's preconceptions. Information that contradicts those preconceptions is either ignored completely or rationalized away. Confirmation bias is problematic because it allows bad decisions to continue long after an impartial weighing of the evidence would have put a stop to them. When you see someone persisting with a failed policy or incorrect opinion despite overwhelming evidence that he or she should try something else, confirmation bias is probably at work.

Self-Serving Bias The term *self-serving bias* refers to people's tendency to attribute their successes to personal effort or personal character traits while at the same time attributing any failures to factors that were out of their control. While helping to preserve people's self-esteem, this bias makes it difficult for people to learn from their mistakes because they incorrectly assume that anything that went wrong was beyond their control.

Overconfidence Effect The *overconfidence effect* refers to people's tendency to be overly confident about how likely their judgments and opinions are to be correct. As an example, people who rated their answers to a particular quiz as being "99 percent likely to be right" were in fact wrong more than 40 percent of the time. Such overconfidence can lead to bad decisions because people will tend to take actions without pausing to verify if their initial hunches are actually true.

Hindsight Bias People engage in *hindsight bias* when they retroactively believe that they were able to predict past events. As an example, consider an election between candidates named Terence and Philip. Before the election happens, many people will predict that Terence will lose. But after Terence ends up winning, many of those same people will convince themselves that they "knew all along" that Terence was going to win. This faulty "I-knew-it-all-along" perspective causes people to massively overestimate their predictive abilities.

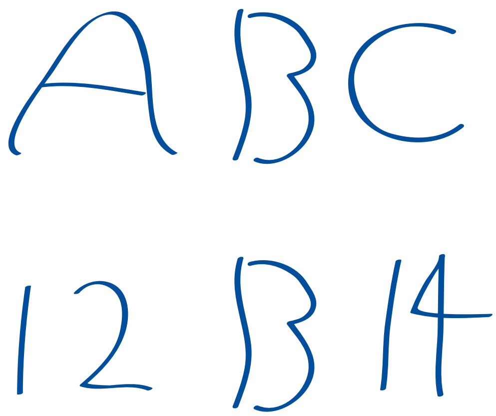
Availability Heuristic The *availability heuristic* causes people to base their estimates about the likelihood of an event not on objective facts but on whether or not similar events come to mind quickly and are readily available in their memories. Because vivid, emotionally charged images come to mind more easily, people tend to think that events like homicides, shark attacks, and lightning strikes are much more common than they actually are. At the same time, they underestimate the likelihood of unmemorable events.

As an example, you are five times more likely to die of stomach cancer than be murdered, but most people rate the likelihood of being murdered as much higher. They do this because they have many vivid memories of both real and fictional murders but almost no recollections whatsoever of anyone dying of stomach cancer.

The availability heuristic causes people to spend too much of their time and effort attempting to protect themselves against charismatic dangers of low actual probability while neglecting to protect themselves against dull threats of substantially higher probability.

Planning Fallacy The *planning fallacy* is the tendency people have to massively underestimate the time needed to

FIGURE 8.3 The letter illusion is the result of a framing effect. In each row, the middle symbol is the same. When that symbol is surrounded by the letters A and C in the top row, our brains tend to register the symbol as the letter B. But when it is surrounded by the numbers 12 and 14 in the bottom row, our brains tend to register it as the number 13. What our brain "sees" is largely a matter of context (frame).



complete a task. A good example is when last-minute test cramming gets really frantic. The student doing the cramming probably underestimated by many hours how much time he needed to prepare for the exam. The planning fallacy also helps to explain why construction projects, business initiatives, and government reform efforts all tend to come in substantially behind schedule.

Framing Effects **Framing effects** occur when a change in context (frame) causes people to react differently to a particular piece of information or to an otherwise identical situation.

Figure 8.3 gives an example of a framing effect. The middle symbol is identical in both rows, but it is interpreted differently depending upon whether it is surrounded by letters or numbers. When surrounded by letters in the top row, the brain tends to interpret the symbol as the letter B. When surrounded by numbers in the bottom row, the brain tends to interpret the symbol as the number 13.

Changes in context can also cause extraordinary changes in behavior. Experiments have shown that ordinary people are twice as likely to litter, steal, or trespass if experimenters tag an area with graffiti and scatter lots of trash around. By changing the area's appearance from neat and orderly to rundown and chaotic, experimenters got ordinary people to subconsciously choose to engage in more crime.

Framing effects can also cause consumers to change their purchases. At the local supermarket, apples command a higher price if each one comes with a pretty sticker and meat sells faster if it is packaged in shiny plastic containers. At a high-end retailer, expensive packaging increases the perceived value of the shop's merchandise. So does having a nice physical space in which to shop. Thus, high-end retailers spend a lot on architecture and displays.

QUICK REVIEW 8.1

- Behavioral economics differs from neoclassical economics because its models of decision-making take into account the fact that heuristics and cognitive biases cause people to make systematic errors.
- To conserve energy, the brain relies on low-energy mental shortcuts, or heuristics, that will usually produce the correct decision or answer.
- Cognitive biases are systematic misperceptions or bad decisions that arise because (1) heuristics are error-prone in certain situations or (2) evolution did not prepare our brains to handle many modern tasks such as solving calculus problems.

Prospect Theory

LO8.3 Relate how prospect theory helps to explain many consumer behaviors, including framing effects, mental accounting, anchoring, loss aversion, and the endowment effect.

Neoclassical economics focuses much of its attention on consumer-choice situations in which people only have to deal with "goods" as opposed to "bads." When deciding on how to spend a budget, a consumer only considers items that would bring her positive marginal utility—that is, "good" things. She then uses the utility-maximizing rule to select how much of each of those good things she should consume to get as much utility as possible from her limited budget.

Unfortunately, life often forces us to deal with bad things, too. Our houses may burn down. A potential investment may go bad. The money we lend out may not be repaid.

How people cope with negative possibilities is a central focus of behavioral economics. Many thousands of observations have been cataloged as to how people actually deal with the prospect of bad things as well as good things. Three very interesting facts summarize how people deal with goods and bads:

- People judge good things and bad things in relative terms, as gains and losses relative to their current situation, or **status quo.**
- People experience both diminishing marginal utility for gains (meaning that each successive unit of gain

feels good, but not as good as the previous unit) as well as diminishing marginal disutility for losses (meaning that each successive unit of loss hurts, but less painfully than the previous unit).

- People experience **loss aversion,** meaning that for losses and gains near the status quo, losses are felt *much* more intensely than gains—in fact, about 2.5 times more intensely. Thus, for instance, the pain experienced by an investor who loses one dollar from his status quo level of wealth will be about 2.5 times more intense than the pleasure he would have felt if he had gained one dollar relative to his status quo level of wealth.

CONSIDER THIS . . .

Rising Consumption and the Hedonic Treadmill

For many sensations, people's brains are wired to notice changes rather than states. For example, your brain can sense acceleration—your change in speed—but not speed itself. As a result, standing still feels the same as moving at a constant 50 miles per hour. And if you accelerate from one constant speed to another—say, from 50 miles per hour to 70 miles per hour—you will feel the acceleration only while it's happening. Once you settle down at the new higher speed, it will feel like you are standing still again.

Consumption appears to work in much the same way. If you are used to a given level of consumption—say, $50,000 per year—then you will get a lot of enjoyment for a while if your consumption accelerates to $100,000 per year. But, as time passes, you will get used to that higher level of consumption, so that $100,000 per year seems ordinary and doesn't bring you any more pleasure than $50,000 per year used to bring you when it was your status quo.

Economist Richard Easterlin coined the term *hedonic treadmill* (pleasure treadmill) to describe this phenomenon. Just as a person walking on a real treadmill gets nowhere, people trying to make themselves permanently happier by consuming more also get nowhere because they end up getting used to any higher level of consumption. Indeed, except for the extremely poor, people across the income spectrum report similar levels of happiness and satisfaction with their lives. This has led several economists, including Robert Frank, to argue that we should all stop trying to consume more because doing so doesn't make us any happier in the long run. What do you think? Should we all step off of the hedonic treadmill?

These three facts about how people deal with goods and bads form the basis of **prospect theory,** which sheds important light on how consumers plan for and deal with life's ups and downs as well as why they often appear narrow-minded and fail to "see the big picture." To give you an idea of how powerful prospect theory is—and why its pioneer, Daniel Kahneman, was awarded the Nobel Prize in Economics—let's go through some examples of consumer behavior that would be hard to explain without the insights provided by prospect theory.

ORIGIN OF THE IDEA

O8.1

Prospect theory

Losses and Shrinking Packages

Because people see the world in terms of gains and losses relative to the status quo situations that they are used to, businesses have to be very careful about increasing the prices they charge for their products. This is because once consumers become used to a given price, they will view any increase in the price as a loss relative to the status quo price they had been accustomed to.

The fact that consumers may view a price increase as a loss explains the otherwise curious fact that many food producers react to rising input costs by shrinking the sizes of their products. The company most famous for doing this is Hershey's chocolates. During its first decades of operation about 100 years ago, it would always charge exactly 5 cents for one of its Hershey's chocolate bars. But the size of the bars would increase or decrease depending on the cost of the company's inputs. When the cost of raw materials rose, the company would keep the price fixed at 5 cents but decrease the size of the bar. When the cost of raw materials fell, it would again keep the price fixed at 5 cents but increase the size of the bar.

This seems rather bizarre when you consider that consumers were not in any way *actually* being shielded from the changes in input prices. That is because what should rationally matter to consumers is the price per ounce that they are paying for Hershey's Bars. And that *does* go up and down when the price remains fixed but the size of the bars changes.

But people aren't being fully rational here. They mentally fixate on the product's price because that is the characteristic that they are used to focusing on when making their purchasing decisions. And because the 5-cent price had become the status quo that they were used to, Hershey's understood that any price increase would be mentally categorized as a loss. Thus, Hershey's wisely chose to keep the price of its product fixed at 5 cents even when input prices were rising.

Other companies employ the same strategy today. In the years following the 2007–2008 recession, the prices of many raw materials, including sugar, soybeans, and corn, rose substantially. Many major manufacturers reacted by reducing product sizes while keeping prices fixed. Häagen-Dazs reduced the size of its supermarket ice cream tubs from 16 to 14 ounces. Kraft reduced the number of slices of cheese in a package of Kraft Singles from 24 to 22 slices. A bottle of Tropicana orange juice shrank from 64 ounces (the traditional half-gallon size) to just 59 ounces. And Procter and Gamble reduced the size of Bounty paper towel rolls from 60 to 52 sheets.

Framing Effects and Advertising

Because people evaluate situations in terms of gains and losses, their decision-making can be very sensitive to the mental frame that they use to evaluate whether a possible outcome should be viewed as a gain or a loss. Here are a couple of examples in which differences in the context or "frame" change the perception of whether a situation should be treated as a gain or loss. See how you react to them.

- Would you be happy with a salary of $100,000 per year? You might say yes. But what if your salary last year had been $140,000? Are you still going to say yes? Now that you know you are taking a $40,000 pay cut, does that $100,000 salary seem as good as it did before?
- Similarly, suppose you have a part-time job. One day, your boss Joe walks in and says that he is going to give you a 10 percent raise. Would that please you? Now, what if he also mentioned that *everyone else* at your firm would be getting a 15 percent raise. Are you still going to be just as pleased? Or does your raise now seem like a loss compared to what everyone else will be getting?

Prospect theory takes into account the fact that people's preferences can change drastically depending on whether contextual information causes them to define a situation as a gain or a loss. These framing effects are important to recognize because they can be manipulated by advertisers, lawyers, and politicians to try to alter people's decisions. For instance, would an advertising company be better off marketing a particular brand of hamburger as "20% fat" or as "80% lean"? Both phrases describe the same meat, but one frames the situation as a loss (20 percent fat) while the other frames it as a gain (80 percent lean).

And would you be more willing to take a particular medicine if you were told that 99.9 percent of the people who take it live or if you were told that 0.1 percent of the people who take it die? Continuing to live is a gain, whereas dying is clearly a loss. Which frame sounds better to you?

Framing effects have major consequences for consumer behavior because any frame that alters whether consumers consider a situation to be a gain or a loss *will* affect their consumption decisions!

Anchoring and Credit Card Bills

Before people can calculate their gains and losses, they must first define the status quo from which to measure those changes. But it turns out that irrelevant information can unconsciously influence people's feelings about the status quo. Here's a striking example. Find a group of people and ask each person to write down the last two digits of his or her Social Security number. Then ask each person to write down his or her best estimate of the value of some object that you display to them—say, a nice wireless keyboard. What you will find is that the people whose Social Security numbers end in higher numbers—say, 67 or 89—will give higher estimates for the value of the keyboard than people whose Social Security numbers end in smaller numbers like 18 or 37. The effect can be huge. Among students in one MBA class at MIT, those with Social Security numbers ending between 80 and 99 gave an average estimate of $56 for a wireless keyboard, while their classmates whose Social Security numbers ended between 00 and 20 gave an average estimate of just $16.

Psychologists and behavioral economists refer to this phenomenon as **anchoring** because people's estimates about the value of the keyboard are influenced, or "anchored," by the recently considered information about the last two digits of their Social Security numbers. Why irrelevant information can anchor subsequent valuations is not fully understood. But the anchoring effect is real and can lead people to unconsciously alter how they evaluate different options.

Unfortunately, credit card companies have figured this out. They use anchoring to increase their profits by showing very small minimum-payment amounts on borrowers' monthly credit card statements. The companies could require larger minimum payments, but the minimum-payment numbers that they present are only typically about 2 percent of what a customer owes. Why such a small amount? Because it acts as an anchor that causes people to unconsciously make smaller payments each month. This can make a huge difference in how long it takes to pay off their bill and how much in total interest they will end up paying. For a customer who owes $1,000 on a credit card that charges the typical interest rate of 19 percent per year, it will take 22 years and $3,398.12 in total payments (including accumulated interest) to pay off the debt if he only makes 2 percent monthly payments. By showing such small minimum-payment amounts, credit card companies anchor many customers into the expensive habit of paying off their debts slowly rather than quickly.

Mental Accounting and Overpriced Warranties

The utility-maximizing rule (Chapter 7) assumes that people will look at all of their potential consumption options simultaneously when trying to maximize the total utility that they can get from spending their limited incomes. But economist Richard Thaler famously noted that people sometimes look at consumption options in isolation, thereby irrationally failing to look at all their options simultaneously. Thaler coined the term **mental accounting** to describe this behavior because it was as if people arbitrarily put certain options into totally separate "mental accounts" that they dealt with without any thought to options outside of those accounts.

As an example of where this suboptimal tendency leads, consider the extended warranties offered by big electronic stores whenever customers purchase expensive products like plasma TVs. These warranties are very much overpriced given that the products they insure hardly ever break down. Personal financial experts universally tell people not to buy them. Yet many people do buy them because they engage in mental accounting.

They do this by mentally labeling their purchase of the TV as an isolated, individual transaction, sticking it into a separate mental account in their brain that might have a title like "Purchase of New TV." Viewing the purchase in isolation exaggerates the size of the potential loss that would come from a broken TV. Customers who view the transaction in isolation see the possibility of a $1,000 loss on their $1,000 purchase as a potential total loss— "Holy cow! I could lose $1,000 on a $1,000 TV!" By contrast, people who can see the big picture are able to compare the potential $1,000 loss with the much larger value of their entire future income stream. That allows them to realize that the potential loss is relatively minor—and thus not a good enough reason to purchase an expensive warranty.

The Endowment Effect and Market Transactions

Prospect theory also offers an explanation for the **endowment effect,** which is the tendency that people have to put a higher valuation on anything that they currently possess (are endowed with) than on identical items that they do not own but might purchase. For instance, if we show a person a new coffee mug and ask him what the maximum amount is that he would pay to buy it, he might say $10. But if we then give the mug to him so that he now owns it, and we then ask how much we would have to pay him to buy it back, he will very likely report a much higher value—say, $15.

The interesting thing is that he is not just bluffing or driving a hard bargain. Rather, the human brain appears wired to put a higher value on things we own than on things we don't. Economist John List has shown that this tendency can moderate if people are used to buying things for resale—that is, buying them with the intention of getting rid of them. But without such experience, the endowment effect can be quite strong. If it is, it can make market transactions between buyers and sellers harder because sellers will be demanding higher prices for the items they are selling ("Hey, *my* mug is worth $15 to me!") than the values put on those items by potential buyers ("Dude, *your* mug is only worth $10 to me").

Several researchers have suggested that loss aversion may be responsible for the endowment effect and the higher values demanded by sellers. They argue that once a person possesses something, the thought of parting with it seems like a potential loss. As a result, the person will demand a lot of money as compensation if he or she is asked to sell the item. On the other hand, potential purchasers do not feel any potential sense of loss, so they end up assigning lower values to the same items.

Status Quo Bias

Prospect theory also explains **status quo bias,** which is the tendency that people have to favor any option that is presented to them as being the default (status quo) option. As an example, consider Global Perspective 8.1. It shows, for a selection of European countries, the percentages of their respective populations that have indicated their willingness to participate in organ-donation programs.

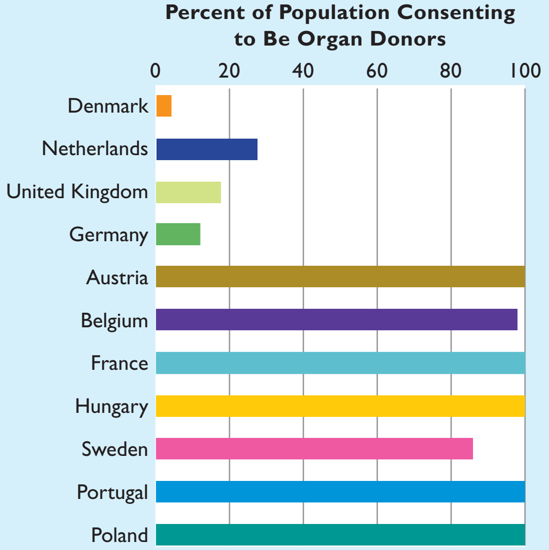
As you can see, seven of the 11 countries have very high participation rates while the other four have low participation rates. You might suspect that cultural differences are at play, but that doesn't make sense when you note that countries like Germany and Austria that are culturally very similar still have massively different participation rates.

What is actually going on is a difference in the default option that people are presented with when they are asked whether they wish to participate. In the seven countries with high participation rates, the default option is participation, so that those who don't want to participate must

GLOBAL PERSPECTIVE 8.1

Percent of Population Consenting to Be Organ Donors

People tend to stick with whatever option is presented as the default option. Thus, the seven countries with high percentages consenting to be organ donors have organ-donation programs in which the default option is participation. By contrast, the four countries with low percentages consenting to be organ donors have organ-donation programs where the default option is *not* participating.



Source: Eric Johnson and Daniel Goldstein, "Defaults and Donation Decisions," *Transplantation* 78, no. 12, December 27, 2004. Used by permission of Wolters Kluwer Health via Copyright Clearance Center.

explicitly check off a box indicating that they don't want to participate. By contrast, in the four countries with low participation rates, the default option is *not* participating, so that those wishing to participate must explicitly check off a box indicating that they want to participate.

What we see in all countries is that nearly everyone chooses to do nothing. They almost never check off the box that would indicate doing the opposite of the default option. Consequently, they end up agreeing to whatever the default option happens to be. Thus, the huge differences in participation rates among the 11 countries are driven almost entirely by what the default option happens to be.

Prospect theory explains this and other examples of status quo bias as a combination of the endowment effect and loss aversion. When people are put into a novel situation, they have no preexisting preferences for any of the options. As a result, the way the options are framed becomes very important because if any of them is presented as the default option, people will tend to treat it as an endowment that they wish to hold on to. At the same time, they will treat any other option as a prospect that could potentially cause a loss. Loss aversion then kicks in and causes most people to stick with the default option in order to avoid the possibility of incurring a loss. The result is a bias toward the status quo.

Status quo bias can be used to explain several consumer behaviors. Consider brand loyalty. If you have gotten used to eating Heinz ketchup, then status quo bias will make you reluctant to purchase any other brand of ketchup. Overcoming that feeling of potential loss is a difficult challenge for competing brands, as attested to by the fact that rivals seeking to challenge an established brand are often forced to resort to deep discounts or free samples to get consumers to even try their products.

Myopia and Time Inconsistency

LO8.4 Describe how time inconsistency and myopia cause people to make suboptimal long-run decisions.

Our ancient ancestors had little cause to spend much time worrying about anything that would happen in the distant future. Infectious diseases, predatory animals, and the constant threat of starvation made life extremely precarious. Consequently, they had to be almost entirely focused on the present moment and how to get through the next few weeks or the next few months.

Today, however, people living in industrialized countries only rarely die from infectious diseases, mostly see predatory animals in zoos, and are under no threat at all of starvation. Living past 80 is now routine and most of us will die of old age. As a result, long-run challenges like planning for retirement and saving for college are now common tasks that nearly everyone faces.

Unfortunately, our brains were designed for our ancestors' more immediate concerns. Thus, we often have difficulty with long-run planning and decisions that involve trade-offs between the present and the future. Two of the major stumbling blocks are myopia and time inconsistency.

Myopia

In biology, myopia, or nearsightedness, refers to a defect of the eye that makes distant objects appear fuzzy, out of focus, and hard to see. By analogy, economists use the word **myopia** to describe the fact that our brains have a hard time conceptualizing the future. Compared with the present, the future seems fuzzy, out of focus, and hard to see.

As an example, our brains are very good at weighing current benefits against current costs in order to make immediate decisions. But our brains almost seem "future blind" when it comes to conceptualizing either future costs or future benefits. As a result, we have difficulty evaluating possibilities that will occur more than a few weeks or months into the future.

The primary consequence of myopia is that when people are forced to choose between something that will generate benefits quickly and something that won't yield benefits for a long time, they will have a very strong tendency to favor the more immediate option. As an example, imagine that Terence has $1,000 that he can either spend on a vacation next month or save for his retirement in 30 years.

Myopia will cause him to have great difficulty imagining the additional spending power that he will be able to enjoy in 30 years if he saves the money. On the other hand, it is very easy for him to imagine all the fun he could have next month if he were to go on vacation. As a result, he will be strongly biased toward spending the money next month. With myopia obscuring the benefits of the long-term option, the short-term option will seem much more attractive.

Myopia also makes it hard to stick with a diet or follow an exercise plan. Compared with the immediate and clearly visible pleasures of eating doughnuts or hanging out, the future benefits from eating better or exercising consistently are just too hazy in most people's minds to be very attractive.

Time Inconsistency

Time inconsistency is the tendency to systematically misjudge at the present time what you will want to do at some future time. This misperception causes a disconnect between what you currently think you will want to do at some particular point in the future and what you actually end up wanting to do when that moment arrives. It is as though your present self does not understand what your future self will want.

Waking up early is a good example. At 8 p.m. on a Tuesday, you may really like the idea of waking up early the next morning so that you can exercise before starting the rest of your day. So you set your alarm 90 minutes earlier than you normally do. But when your alarm goes off the next morning at that earlier time, you loath the concept, throw the alarm across the room, and go back to sleep. That switch in your preferences from the night before is the essence of time inconsistency. Your future self ends up disagreeing with your current self.

Self-Control Problems Time inconsistency is important because it is a major cause of **self-control problems.** To see why, imagine that before heading out to a restaurant with friends, you think that you will be happy sticking to your diet and only ordering a salad. After all, that particular restaurant has very tasty salads. But then, after you get there, you find the dessert menu overwhelmingly attractive and end up ordering two servings of cheesecake.

Because you were time inconsistent and didn't understand what your future self would want, you placed yourself into a situation in which it was very difficult for you to stick to your diet. If you had, instead, been able to correctly predict what your future self would want, you might have decided to stay home for the evening rather than putting yourself in temptation's way. Alternatively, you could have gone, but not before making your friends promise to prevent you from ordering dessert.

Time inconsistency also makes it hard for many workers to save money. Before their paychecks arrive, they mistakenly assume that their future selves will want to save money as much as their current selves do. But once the money becomes available, their future selves end up wanting to spend everything and save nothing.

Fighting Self-Control Problems with Precommitments The key to fighting time inconsistency and self-control problems is to have a good understanding of what your future self is likely to want. You can then make **precommitments** and take actions ahead of time to prevent your future self from doing much damage.

Hiding the Alarm Clock Consider again the problem of wanting to wake up 90 minutes early on Wednesday morning so that you can work out before starting the rest of your day. If you understand that your future self is not going to want to cooperate, you can take steps to prevent that future self from flaking out. Some people set multiple alarms. Others put their alarms on the other side of the room, underneath a pile of stuff that will have to be moved if the future self wants to turn the damned thing off. But the point is that each of these methods ensures that it will be nearly impossible for the future self to easily get back to sleep. They set things up so that the future self will be forced to do what the present self desires.

Automatic Payroll Deductions Precommitment strategies have also been used to help future selves save more. Consider automatic payroll deductions. If a worker named

Blaire signs up for such a program, a fixed percentage will be automatically deducted from each of her paychecks and deposited directly into her retirement savings account. Because that money never gets to her checking account, there is no way for Blaire's future self to fall prey to temptation and spend it. As the old saying goes, "Out of sight, out of mind."

Salary Smoothing School teachers and college professors often have the choice of having their annual salaries paid out over 9 larger monthly installments (to match the length of the school year) or 12 smaller monthly installments (to match the length of the calendar year). If we observe which option they actually choose, we find that the vast majority opt to have their salaries spread out over 12 months rather than 9 months.

They do so because they fear self-control problems. In particular, they are afraid that if they opt to be paid over 9 months, they won't have the self control to save enough money during the 9-month period when they will be getting paid to last them through the three months of summer vacation when they won't be getting paid. To avoid that situation, they opt to have their salaries spread out evenly over the entire calendar year. That precommitment ensures that their future selves are never given the chance to blow through all the money too quickly.

Early Withdrawal Penalties Sometimes, one cognitive bias can be used to offset another. Retirement accounts that have early-withdrawal penalties are a good example. They use loss aversion to offset time inconsistency and self-control problems.

In some cases, the penalties on these sorts of accounts are as high as 25 percent, meaning that if a saver wanted to withdraw $1,000 before reaching retirement, he would have to give up an additional $250 (= 25 percent of $1,000) as a penalty. While that amount is substantial in itself, loss aversion makes it even more painful to contemplate. As a result, most people can't bring themselves to make an early withdrawal.

Weight-Loss Competitions Loss aversion also drives the effectiveness of weight-loss competitions. For a person who has agreed to participate, the prospect of losing the competition can be a great motivator because loss aversion applies just as much to future selves as to present selves. Even after the future rolls around and the future self is in charge, the future self won't like the prospect of losing either. Thus, the present self can be confident that the future self will also be motivated to stick to the weight-loss goals that the present self wants to achieve.

CONSIDER THIS . . .

Betting Against Yourself

The website StickK.com makes it easy for people to set up financial incentives that can help them reach their goals. Strangely enough, it does that by getting people to bet against themselves.

Founded by behavioral economist Dean Karlan and fellow Yale University professor Alan Iyers, StickK.com lets each user specify a goal, such as losing 30 pounds in the next year. The user then sets up a "commitment contract" that specifies an amount of money that he or she will have to pay via credit card as a penalty for not reaching his or her goal. That potential penalty gets loss aversion to kick in, thereby helping the user follow through and achieve his or her goal.

As a fun twist, StickK.com also requires each user to specify a recipient for the penalty money. Some users specify that their penalty money go to a charity that they like. But many choose to increase their motivation by designating an "anti-charity" as the recipient. What is an anti-charity? It's a person or group that the user hates.

By letting people designate anti-charities, StickK.com uses deeply seated personal animosities in addition to loss aversion to help people overcome time inconsistency and self-control issues. Those incentives appear to work quite well, as 70 percent of StikK.com users report success in reaching their goals.

The nearby Consider This piece illustrates another effective way to use loss aversion to help overcome self-control problems.

QUICK REVIEW 8.2

- Prospect theory models decision-making by accounting for the fact that people's choices are affected by whether a possible outcome is perceived as a prospective gain or a prospective loss relative to the current status quo situation.
- Because our ancestors were focused on short-term survival, our brains suffer from myopia and are not good at dealing with decisions that involve the future.
- Precommitments can be used to compensate for time inconsistency and the self-control problems that arise when the future self doesn't want to do what the present self prefers.

Fairness and Self-Interest

LO8.5 Define fairness and give examples of how it affects behavior in the economy and in the dictator and ultimatum games.

Neoclassical models assume that people are purely self-interested. They do so because "pure self-interest" seems like a good basis for predicting many economic behaviors, especially those happening in market situations where people are dealing mostly with strangers and are, consequently, unlikely to be particularly sentimental or charity-minded.

Adam Smith, the founder of modern economics, put this line of thinking into words. The most-quoted passage from *The Wealth of Nations* reads,

> It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves not to their humanity but to their self-love, and never talk to them of our own necessities but of *their* advantages.

Smith, however, did not believe that people are *exclusively* focused on self-love and their own interests. He believed that we are also strongly motivated by emotions such as charity, selflessness, and the desire to work for the common good. He expressed this view at length in his other influential book, *The Theory of Moral Sentiments*. The most-famous passage from that book reads:

> How selfish soever man may be supposed, there are evidently some principles in his nature which interest him in the fortune of others and render their happiness necessary to him though he derives nothing from it except the pleasure of seeing it.

What behavioral economists have discovered is that this human propensity to care about others extends into every type of economic behavior. While self-interest is always present, most people care deeply about others and how they are interacting with others. As a result, economic transactions are heavily influenced by moral and ethical factors.

Field Evidence for Fairness

Many real-world behaviors support the contention that economic transactions are heavily influenced by beliefs and values. This "field evidence" has helped behavioral economists identify the ethical and moral factors that appear to have the largest influence on economic behavior. Fairness is among the most important.

Fairness is a person's opinion as to whether a price, wage, or allocation is considered morally or ethically acceptable. Standards of fairness vary from person to person and economists generally take no stand on what people consider to be right or wrong. But fairness has been studied extensively because many everyday economic behaviors indicate that people care substantially about fairness and not just about maximizing what they can get for themselves.

Consider the following examples—none of which would be undertaken by a purely self-interested person.

- ***Giving to Charity*** Each year, U.S. charities receive over $300 billion of cash donations and 8 billion hours of free labor. These donations of time and money are inconsistent with the idea that people are only interested in themselves. What is more, many of the cash donations are anonymous. That suggests that many donors have extremely pure motives and are not donating just to make themselves look good.
- ***Obeying the Law*** In many countries, the large majority of citizens are law-abiding despite having many opportunities to break the law without getting caught. In the same way, the large majority of taxpayers complete their tax returns honestly despite having many opportunities to cut corners and hide income.
- ***Fixing Prices*** During hurricanes and other natural disasters, shortages of crucial products such as gasoline and electric generators often develop. The shortages imply that retailers could raise prices, but they mostly keep prices fixed because they do not want to be thought of as taking advantage of the situation.
- ***Purchasing "Fair-Trade" Products*** Many consumers are willing to pay premium prices to purchase products that have been certified by the Fair Trade organization as having been produced by companies that meet high standards with respect to workers' rights and environmental sustainability. These customers clearly care about more than just getting the lowest price.

Experimental Evidence for Fairness

Our understanding of fairness and how it affects economic transactions has been reinforced and refined in recent decades by examining experimental games that were specifically designed to test people's feelings about fairness.

The most important feature of these games is that they are played for real money. That matters because if people were only motivated by self-interest, you would expect everyone playing the games to utilize only those strategies that are most likely to maximize their own winnings.

LAST WORD

Nudging People Toward Better Decisions

Behavioral Economists Have Recently Found Success in Using People's Behavioral Biases to "Nudge" Them Toward Making Better Decisions.*

Behavioral economics began as a descriptive science, meaning that its first goal was to develop theories that accurately described human economic behavior. In particular, it sought to explain a number of behaviors that at first glance seemed irrational. Now that behavioral economics has made significant headway in explaining many of those behaviors, some economists are suggesting that its insights be used to nudge people toward choices that are better for themselves and others.

A key feature of "nudges" is that they are subtle. This subtlety means that nudges can cause large changes in behavior without making people feel bullied or coerced—and also without imposing stringent new rules or having to offer people big monetary incentives or disincentives to get them to do what you want.

*The term "nudge" was popularized by Richard Thaler and Cass Sunstein in their book *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008.

Take retirement savings. Myopia and time inconsistency cause many people to consume too much in the present and therefore undersave for retirement. But as it turns out, this unfortunate behavioral tendency can be easily offset by utilizing status quo bias and people's tendency to stick with default options. In terms of retirement savings, this comes down to designing corporate retirement programs in which each worker is "defaulted into" her company's retirement savings program.

Under those savings programs, money is automatically deducted each month from a worker's paycheck and deposited in her retirement savings account. It used to be the case that the default for such programs was for workers to start out *not* enrolled in them. To get enrolled, they would have to request to join the program. That is, they would have to choose to go against the default option of not being enrolled.

As it turns out, however, only a few people behave that way. The majority actually play fairly and generously, often going out of their way to share with less-fortunate players even when they are under no compulsion to do so. That being said, their kindness only goes so far. If other players are acting selfishly, the average person will withhold cooperation and may even retaliate.

The Dictator Game The strongest experimental evidence against the idea that people are only interested in what they can get for themselves comes from the **dictator game.**

The Rules In the game, two people interact anonymously. One of them is randomly designated as the "dictator." It is his job to split an amount of money that is put up for that purpose by the researcher running the game. A typical amount is $10.

The defining feature of the game is that the dictator can dictate whatever split he prefers. It could be to keep all the money for himself. It could be to give all the money to the other player. It could be to split it in any other possible way, such as $8.67 for himself and $1.33 for the other person.

Because the game is fully anonymous, the dictator doesn't have to worry about retaliation by the other person. He can get away with being as selfish as he wants.

How Players Behave So what actually happens when people play the dictator game? After running the experiment many thousands of times in many different countries, experimenters have found that only one-third of dictators keep all of the money for themselves. The other two-thirds show substantial generosity, allocating an average of 42 percent of the money to the other player. In addition, 17 percent of all dictators split the money perfectly evenly and a little over five percent of all dictators give the other player everything.

Implications for Fairness The way dictators behave suggests two important things about fairness.

First, the majority of people appear to be genuinely concerned about being fair to other people. They are willing to take less for themselves in order to ensure that the other

And because people have the behavioral tendency of sticking with whatever option is presented to them as the default, relatively few workers would make the change and enroll in their company's savings program. That was disappointing. But instead of being deterred, behavioral economists saw an opportunity. Why not change the default? Why not make automatic enrollment the default option? By making that change, people's tendency to stick with default options would work in their own favor—they would stay enrolled and save money for retirement.

When this strategy of switching the default was actually implemented, the number of workers participating in retirement savings programs skyrocketed—jumping from 60 percent to 98 percent. Those workers can now look forward to much more pleasant retirements thanks to this simple change that works *with* people's preference to stick with default options.

People's tendency to look around them for social cues as to what constitutes good behavior can also be exploited to modify their consumption behavior. But you have to be careful about how you do it, as was discovered by Opower, an energy consulting firm that wanted to encourage customers to conserve electricity. Its first attempt to use social cues involved sending each customer in a California town a bill that showed not only his or her own usage of electricity in kilowatt-hours, but also the average usage of nearby houses. Opower hoped that by showing the average usage of neighbors, customers would receive a subtle hint about their own usage. In particular, it was hoped that customers who used more than their neighbors would feel that they were being wasteful and would thus cut back on their usage.

And that did indeed happen. *But* their reduction in electricity usage ended up being completely swamped by an increase in electricity usage on the part of the customers who had previously been below-average users. Those customers interpreted the new information that they were below-average electricity users to mean that they should feel free to consume more. After all, why should they use so little when their neighbors were using so much more?

Taking that into account, Opower finally hit upon a solution that worked. Smilies. Yes, symbols like ☺ and ☹. In addition to printing people's own usage and the average usage of their neighbors, Opower also started printing a ☺ on a customer's bill if his usage was below average and a ☹ on his bill if his usage was above average. The unhappy smilies embarrassed the heavy users into reducing their consumption even more, while the happy smilies gave a pat on the back to the light users—a pat on the back that kept their usage low.

Bear in mind that both the electricity customers and the workers saving for retirement were being *manipulated* by the people who designed the nudges. This fact is perhaps even more disturbing when you consider that the changes in behavior that were caused by the nudges were most likely *unconscious* on the part of those being manipulated. Keep this in mind as you consider for yourself when and if it is morally or ethically acceptable to use nudges to guide people's behavior.

player receives something, too. And they are willing to give substantially to the other player even though the game's guarantee of anonymity would allow them to take everything for themselves without fear of retaliation.

Second, generosity varies quite widely. Between the third of dictators who keep everything for themselves and the five percent who give everything to the other person lie the large majority who allocate some but not all of the money to the other person. Within that group, every possible split of the money can be found. As a result, behavioral economists believe that individuals vary widely in their beliefs about fairness. Some are incredibly selfish. Others are incredibly generous. And most of us lie somewhere in between.

To help get a better handle on how those widely divergent beliefs affect behavior in more realistic situations, economists designed a slightly more complex game.

The Ultimatum Game Like the dictator game, the **ultimatum game** involves two players anonymously splitting an amount of money. But there is no longer a dictator who can arbitrarily decide how the money is spilt. Instead, both players need to agree on any proposed split if it is to take place.

That difference in the rules ensures that the ultimatum game mirrors the many real-world situations in which a project or proposal must obtain the consent and support of all parties if it is to be undertaken. As an example, consider a business transaction between a potential seller and a potential buyer. Even if there are substantial net benefits available to both parties, no transaction will take place unless the buyer and the seller can come to an agreement on the selling price.

The Rules As with the dictator game, the researcher puts up an amount of money to be split. This pot of money is similar in spirit to the net benefits that a buyer and a seller can split if they can agree on a price. It also represents the net benefits that will be forgone if the two parties cannot reach an agreement.

At the start of the experiment, one of the players is randomly assigned to be "the proposer" while the other player is randomly assigned to be "the responder." The game begins with the proposer proposing a split. As in the dictator game, the proposed split can range anywhere from suggesting that all the money go to the proposer to suggesting that all the money go to the responder.

The responder examines the proposed split and decides whether to accept it or reject it. If she accepts it, the split is made and both players are immediately paid their shares by the researcher. But if the responder rejects the proposed split, neither player gets anything. The game simply ends and both players go home without receiving any money at all—a situation similar to when a business negotiation fails and all the potential benefits are forgone.

How Players Behave When the ultimatum game is played, two behaviors stand out.

The more important is that the splits proposed by proposers in the ultimatum game are much more equal on average than the splits imposed by dictators in the dictator game. This is best seen by noting that whereas one-third of dictators keep all the money for themselves in the dictator game, almost no proposers suggest allocating all the money to themselves in the ultimatum game.

This extremely large difference in behavior arises because the people acting as proposers in the ultimatum game realize that suggesting a highly unequal split is almost certain to greatly offend a responder's sense of fairness and lead to a rejection. In addition, most proposers also seem to understand that even moderately unfair offers might also offend responders. As a result, the large majority of proposers suggest either perfectly equal splits or splits that are only slightly biased in the proposer's favor (such as 55 percent going to the proposer).

The second behavior that stands out is the decisiveness and emotional intensity with which responders reject offers that they consider unfair. Of particular interest is the fact that rejection decisions are not made in a cool and calculating fashion. Responders do *not* calmly weigh the costs and benefits of accepting an unfair offer. They actually become extremely angry and reject as a way of retaliating against the proposer. Their rejections are not just negative responses; they are acts of vengeance designed to hurt the proposer by denying him money.

The full extent to which unfair offers make responders angry can be gauged by looking at high-stakes versions of the ultimatum game in which proposers and responders attempt to split hundreds or even thousands of dollars. You might think that when such large amounts of money are on the line, responders would be willing to accept unfair splits. But what we actually see is responders continuing to reject splits that they consider to be unfair. Their preference for fair treatment is so strong that they will reject unfair offers even when doing so means giving up a *lot* of money.

Why the Threat of Rejection Increases Cooperation Some people won't offer anything to other people unless they are coerced into doing so. This is best understood by comparing the behavior of dictators in the dictator game with the behavior of proposers in the ultimatum game. In the dictator game, a full third of dictators award themselves all the money and leave nothing for the other player. In the ultimatum game, by contrast, nearly every proposer offers a substantial split to the responder.

That dramatic increase in generosity and fairness is, of course, related to the different rules used in the two games. When one person has total control over the split, selfish tendencies are given free reign. But when rejections become possible, the player in charge of proposing the split has to take the other player's feelings into account. That causes even selfish proposers to make generous offers because they quickly realize that the only way they can get any money for themselves is by making proposals that will not be rejected.

Implications for Market Efficiency The willingness of proposers to make more generous offers when faced with the threat of rejection can be thought of as the simplest expression of the invisible hand.

As we discussed in Chapter 2, the invisible hand is a metaphor that summarizes the tendency of the market system to align private interests with social interests and get people behaving in ways that benefit not only themselves but other people, too.

In the case of the ultimatum game, the threat of rejection helps to align private interests with social interests. It does so by motivating selfish people to make substantially more generous offers. The result is a higher level of cooperation and utility as offers get accepted and players split the money.

A similar process can be seen in the real world with respect to consumer sovereignty. As discussed in Chapter 2, consumer sovereignty is the right of consumers to spend their incomes on the goods and services that they are most willing and able to buy. Crucially, that right includes the ability to reject any product that does not meet the consumer's expectations.

That right of rejection leads to substantial social benefits because it motivates producers to work hard at producing products that will be acceptable to consumers. Over time, those efforts lead to increased allocative and productive efficiency as better products get produced at lower prices.

SUMMARY

LO8.1 Define behavioral economics and explain how it contrasts with neoclassical economics.

Neoclassical economics bases its predictions about human behavior on the assumption that people are fully rational decision-makers who have no trouble making mental calculations and no problems dealing with temptation. While some of its predictions are accurate, many are not.

The key difficulty facing neoclassical economics is that people make systematic errors, meaning that they regularly and repeatedly engage in behaviors that reduce their likelihood of achieving what they want.

Behavioral economics attempts to explain systematic errors by combining insights from economics, psychology, and biology. Its goal is to make more accurate predictions about human choice behavior by taking into account the mental mistakes that lead to systematic errors.

LO8.2 Discuss the evidence for the brain being modular, computationally restricted, reliant on heuristics, and prone to various forms of cognitive error.

Our brains make systematic errors for two reasons. First, our brains were not prepared by evolution for dealing with many modern problems, especially those having to do with math, physics, and statistics. Second, our brains also make mistakes when dealing with long-standing challenges (like interpreting visual information) because caloric limitations forced our brains to adopt low-energy heuristics (shortcuts) for completing mental tasks.

Heuristics sacrifice accuracy for speed and low energy usage. In most cases, the lack of accuracy is not important because the errors that result are relatively minor. However, in some cases, those errors can generate cognitive biases that substantially impede rational decision-making. Examples include confirmation bias, the overconfidence effect, the availability heuristic, and framing effects.

LO8.3 Relate how prospect theory helps to explain many consumer behaviors, including framing effects, mental accounting, anchoring, loss aversion, and the endowment effect.

Prospect theory is the behavioral economics theory that attempts to accurately describe how people deal with risk and uncertainty. Its key feature is that it models a person's preferences about uncertain outcomes as being based on whether those outcomes will cause gains or losses relative to the current status quo situation to which the person has become accustomed.

Prospect theory also accounts for loss aversion and the fact that most people perceive the pain of losing a given amount of money as being about 2.5 times more intense than the pleasure they would receive from an equal-sized gain.

LO8.4 Describe how time inconsistency and myopia cause people to make suboptimal long-run decisions.

Myopia refers to the difficulty that most people have in conceptualizing the future. It causes people to put insufficient weight on future outcomes when making decisions.

Time inconsistency refers to the difficulty that most people have in correctly predicting what their future selves will want. It causes self-control problems because people are not able to correctly anticipate the degree to which their future selves may fall prey to various sorts of temptation.

People sometimes utilize precommitments to help them overcome self-control problems. Precommitments are courses of action that would be very difficult for the future self to alter. They consequently force the future self to do what the present self desires.

LO8.5 Define fairness and give examples of how it affects behavior in the economy and in the dictator and ultimatum games.

Behavioral economists have found extensive evidence that people are *not* purely self-interested. Rather, they care substantially about fairness and are often willing to give up money and other possessions in order to benefit other people.

The field evidence for fairness includes donations to charity, law-abiding behavior, the reluctance of retailers to raise prices during natural disasters, and the willingness of many consumers to pay premium prices for Fair Trade products.

The dictator and ultimatum games provide experimental evidence on fairness by showing how pairs of people interact to split a pot of money that is provided by the researcher. In the dictator game, one person has total control over the split. In the ultimatum game, both players must agree to the split.

The dictator game shows that many people will share with others even when anonymity would allow them to be perfectly selfish and keep all the money for themselves. The ultimatum game shows that people put a very high value on being treated fairly. They would rather reject an unfair offer and get nothing than accept it and get something.

TERMS AND CONCEPTS

neoclassical economics
behavioral economics
rational
systematic errors
heuristics
cognitive biases

framing effects	mental accounting	self-control problems
status quo	endowment effect	precommitments
loss aversion	status quo bias	fairness
prospect theory	myopia	dictator game
anchoring	time inconsistency	ultimatum game

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Suppose that Joe enjoys and repeatedly does stupid things like getting heavily into debt and insulting police officers. Do these actions constitute systematic errors? If he gets what he wants each time, are his stupid actions even considered to be errors by economists? Explain. **LO8.1**
2. Why do behavioral economists consider it helpful to base a theory of economic behavior on the actual mental processes that people use to make decisions? Why do neoclassical economists not care about whether a theory incorporates those actual mental processes? **LO8.1**
3. Economist Gerd Gigerenzer characterizes heuristics as "fast and frugal" ways of reaching decisions. Are there any costs to heuristics being "fast and frugal"? Explain and give an example of how a fast and frugal method for doing something in everyday life comes at some costs in terms of other attributes forgone. **LO8.2**
4. "There's no such thing as bad publicity." Evaluate this statement in terms of the recognition heuristic. **LO8.2**
5. For each of the following cognitive biases, come up with at least one example from your own life. **LO8.2**
   a. Confirmation bias.
   b. Self-serving bias.
   c. The overconfidence effect.
   d. Hindsight bias.
   e. The availability heuristic.
   f. The planning fallacy.
   g. Framing effects.
6. Suppose that Ike is loss averse. In the morning, Ike's stockbroker calls to tell him that he has gained $1,000 on his stock portfolio. In the evening, his accountant calls to tell him that he owes an extra $1,000 in taxes. At the end of the day, does Ike feel emotionally neutral since the dollar value of the gain in his stock portfolio exactly offsets the amount of extra taxes he has to pay? Explain. **LO8.3**
7. You just accepted a campus job helping to raise money for your school's athletic program. You are told to draft a fundraising letter. The bottom of the letter asks recipients to write down a donation amount. If you want to raise as much money as possible, would it be better if the text of that section mentioned that your school is ranked third in the nation in sports or that you are better than 99 percent of other schools at sports? Explain. **LO8.3**
8. In the early 1990s, New Jersey and Pennsylvania both reformed their automobile insurance systems so that citizens could opt for either a less-expensive policy that did not allow people to sue if they got into accidents or a more-expensive policy that did allow people to sue if they got into accidents. In New Jersey, the default option was the less-expensive policy that did not allow suing. In Pennsylvania, the default option was the more-expensive policy that did allow suing. Given those options, which policy do you think most people in New Jersey ended up with? What about in Pennsylvania? Explain. **LO8.3**
9. Give an example from your own life of a situation where you or someone you know uses a precommitment to overcome a self-control problem. Describe why the precommitment is useful and what it compensates for. Avoid any precommitment that was mentioned in the book. **LO8.4**
10. What does behavioral economics have to say about each of the following statements? **LO8.5**
    a. "Nobody is truly charitable—they just give money to show off."
    b. "America has a ruthless capitalist system. Considerations of fairness are totally ignored."
    c. "Selfish people always get ahead. It's like nobody even notices!"
11. Do people playing the dictator game show only self-interested behavior? How much divergence is there in the splits given by dictators to the other player? **LO8.5**
12. Evaluate the following statement. "We shouldn't generalize from what people do in the ultimatum game because $10 is a trivial amount of money. When larger amounts of money are on the line, people will act differently." **LO8.5**
13. **LAST WORD** What do you think of the ethics of using unconscious nudges to alter people's behavior? Before you answer, consider the following argument made by economists Richard Thaler and Cass Sunstein, who favor the use of nudges. They argue that in most situations, we couldn't avoid nudging even if we wanted to because whatever policy we choose will contain some set of unconscious nudges and incentives that will influence people. Thus, they say, we might as well choose the wisest set of nudges.

REVIEW QUESTIONS

1. Which of the following are systematic errors? **LO8.1**
   a. A colorblind person who repeatedly runs red lights.
   b. An accountant whose occasional math errors are sometimes on the high side and sometimes on the low side.
   c. The tendency many people have to see faces in clouds.
   d. Miranda paying good money for a nice-looking apple that turns out to be rotten inside.
   e. Elvis always wanting to save more but then spending his whole paycheck, month after month.
2. Identify each statement as being associated with neoclassical economics or behavioral economics. **LO8.1**
   a. People are eager and accurate calculators.
   b. People are often selfless and generous.
   c. People have no trouble resisting temptation.
   d. People place insufficient weight on future events and outcomes.
   e. People only treat others well if doing so will get them something they want.
3. Label each of the following behaviors with the correct bias or heuristic. **LO8.3**
   a. Your uncle says that he knew all along that the stock market was going to crash in 2008.
   b. When Fred does well at work, he credits his intelligence. When anything goes wrong, he blames his secretary.
   c. Ellen thinks that being struck dead by lightning is much more likely than dying from an accidental fall at home.
   d. The sales of a TV that is priced at $999 rise after another very similar TV priced at $1,300 is placed next to it at the store.
   e. The sales of a brand of toothpaste rise after new TV commercials announce that the brand "is preferred by 4 out of 5 dentists."
4. Erik wants to save more, but whenever a paycheck arrives, he ends up spending everything. One way to help him overcome this tendency would be to: **LO8.4**
   a. Teach him about time inconsistency.
   b. Tell him that self-control problems are common.
   c. Have him engage in precommitments that will make it difficult for his future self to overspend.
5. Many proposers in the ultimatum game offer half to the responder with whom they are paired. This behavior could be motivated by (select as many as might apply): **LO8.5**
   a. Fear that an unequal split might be rejected by a fair-minded responder.
   b. A desire to induce the responder to reject the offer.
   c. A strong sense of fairness on the part of the proposers.
   d. Unrestrained greed on the part of the proposers.

PROBLEMS

1. One type of systematic error arises because people tend to think of benefits in percentage terms rather than in absolute dollar amounts. As an example, Samir is willing to drive 20 minutes out of his way to save $4 on a grocery item that costs $10 at a local market. But he is unwilling to drive 20 minutes out of his way to save $10 on a laptop that costs $400 at a local store. In percentage terms, how big is the savings on the grocery item? On the laptop? In absolute terms, how big is the savings on the grocery item? On the laptop? If Samir is willing to sacrifice 20 minutes of his time to save $4 in one case, shouldn't he also be willing to sacrifice 20 minutes of his time to save $10? **LO8.2**
2. Anne is a bargain-minded shopper. Normally, her favorite toothpaste costs the same at both of her local supermarkets, but the stores are having competing sales this week. At one store, there is a bonus offer: buy 2, get 1 free. At the other store, toothpaste is being sold at 40 percent off. Anne instantly opts for the first offer. Was that really the less-expensive choice? (Hint: Is "buy 2, get 1 free" the same as 50 percent off?) **LO8.2**
3. The coffee shop near the local college normally sells 10 ounces of roasted coffee beans for $10. But the shop sometimes puts the beans on sale. During some sales, it offers "33 percent more for free." Other weeks, it takes "33 percent off" the normal price. After reviewing the shop's sales data, the shop's manager finds that "33 percent more for free" sells a lot more coffee than "33 percent off." Are the store's customers making a systematic error? Which is actually the better deal? **LO8.2**
4. Angela owes $500 on a credit card and $2,000 on a student loan. The credit card has a 15 percent annual interest rate and the student loan has a 7 percent annual interest rate. Her sense of loss aversion makes her more anxious about the larger loan. As a result, she plans to pay it off first—despite the fact that professional financial advisors always tell people to pay off their highest-interest-rate loans first. Suppose Angela has only $500 at the present time to help pay down her loans and that this $500 will be the only money she will have for making debt payments for at least the next year. If she uses the $500 to pay off the credit card, how much interest will accrue on the other loan over the coming year? On the other hand, if she uses the $500 to pay off part of the student loan, how much in combined interest will she owe over the next year on the remaining balances on the two loans? By how many dollars will she be better off if she uses the $500 to completely pay off the credit card rather than partly paying down the student loan? (Hint: If you owe X dollars at an annual interest rate of Y percent, your annual interest payment will be $X \times Y$, where the interest rate Y is expressed as a decimal.) **LO8.3**

5. **ADVANCED ANALYSIS** In the algebraic version of prospect theory, the variable x represents gains and losses. A positive value for x is a gain, a negative value for x is a loss, and a zero value for x represents remaining at the status quo. The so-called value function, $v(x)$, has separate equations for translating gains and losses into, respectively, positive values (utility) and negative values (disutility). The gain or loss is typically measured in dollars while the resulting value (utility or disutility) is measured in utils. A typical person values gains ($x > 0$) using the function $v(x) = x^{0.88}$ and losses ($x < 0$) using the function $v(x) = -2.5*(-x)^{0.88}$. In addition, if she stays at the status quo ($x = 0$), then $v(x) = 0$. First use a scientific calculator (or a spreadsheet program) and the typical person's value functions for gains and losses to fill out the missing spaces in the nearby table. Then answer the questions that follow. **LO8.3**

Gain or Loss	Total Value of Gain or Loss	Marginal Value of Gain or Loss
−3	−6.57	
−2		−2.10
−1	−2.50	−2.50
0	0.00	—
1		1.00
2	1.84	
3		0.79

a. What is the total value of gaining \$1? Of gaining \$2?

b. What is the marginal value of going from \$0 to gaining \$1? Of going from gaining \$1 to gaining \$2? Does the typical person experience diminishing marginal utility from gains?

c. What is the marginal value of going from \$0 to losing \$1? Of going from losing \$1 to losing \$2? Does the typical person experience diminishing marginal disutility from losses?

d. Suppose that a person simultaneously gains \$1 from one source and loses \$1 from another source. What is the person's total utility after summing the values from these two events? Can a *combination* of events that leaves a person with the same wealth as they started with be perceived negatively? Does this shed light on status quo bias?

e. Suppose that an investor has one investment that gains \$2 while another investment simultaneously loses \$1. What is the person's total utility after summing the values from these two events? Will an investor need to have gains that are bigger than her losses just to feel as good as she would if she did not invest at all and simply remained at the status quo?

6. Ted has always had difficulty saving money, so on June 1, Ted enrolls in a Christmas savings program at his local bank and deposits \$750. That money is totally locked away until December 1 so that Ted can be certain that he will still have it once the holiday shopping season begins. Suppose that the annual rate of interest is 10 percent on ordinary savings accounts (that allow depositors to withdraw their money at any time). How much interest is Ted giving up by precommitting his money into the Christmas savings account for six months instead of depositing it into an ordinary savings account? (Hint: If you invest X dollars at an annual interest rate of Y percent, you will receive interest equal to $X \times Y$, where the interest rate Y is expressed as a decimal.) **LO8.4**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **FOUR**

MICROECONOMICS OF PRODUCT MARKETS

CHAPTER **9**

Businesses and the Costs of Production

Learning Objectives

LO9.1 **Explain why economic costs include both explicit (revealed and expressed) costs and implicit (present but not obvious) costs.**

LO9.2 **Relate the law of diminishing returns to a firm's short-run production costs.**

LO9.3 **Describe the distinctions between fixed and variable costs and among total, average, and marginal costs.**

LO9.4 **Use economies of scale to link a firm's size and its average costs in the long run.**

LO9.5 **Give business examples of short-run costs, economies of scale, and minimum efficient scale (MES).**

Our attention now turns from the behavior of consumers to the behavior of producers. In market economies, a wide variety of businesses produce an even wider variety of goods and services. Each of those businesses requires economic resources in order to produce its products. In obtaining and using resources, a firm makes monetary payments to resource owners (for example, workers) and incurs opportunity costs when using resources it already owns (for example, entrepreneurial talent). Those payments and opportunity costs together make up the firm's *costs of production,* which we discuss in this chapter.

Then, in the next several chapters, we bring product demand, product prices, and revenue back into the analysis and explain how firms compare revenues and costs in determining how much to produce. Our ultimate purpose is to show how those comparisons relate to economic efficiency.

Economic Costs

LO9.1 Explain why economic costs include both explicit (revealed and expressed) costs and implicit (present but not obvious) costs.

Firms face costs because the resources they need to produce their products are scarce and have alternative uses. Because of scarcity, firms wanting a particular resource have to bid it away from other firms. That process is costly for firms because it requires a payment to the resource owner. This reality causes economists to define an **economic cost** as the payment that must be made to obtain and retain the services of a resource. It is the income the firm must provide to resource suppliers to attract resources away from alternative uses.

This section explains how firms incorporate opportunity costs to calculate economic costs. If you need a refresher on opportunity costs, a brief review of the section on opportunity costs in Chapter 1 might be useful before continuing on with the rest of this section.

Explicit and Implicit Costs

To properly calculate a firm's economic costs, you must remember that each of the resources used by the firm has an opportunity cost. This is true both for the resources that a firm purchases from outsiders as well as for the resources that it already owns.

As an example, consider a table-making firm that starts this month with $5,000 in cash as well as ownership of a small oak forest from which it gets the oak that it turns into tables.

Suppose that during the month the firm uses the entire $5,000 of cash to pay its workers. Clearly, the $5,000 it spends purchasing their labor comes at the opportunity cost of forgoing the best alternatives that could have been bought with that money.

Less obvious, however, is the opportunity cost of the oak that the firm grows itself and which it uses to make tables. Suppose that the oak has a market value of $1,500, meaning that our table-making firm could sell it to outsiders for $1,500. This implies that using the oak to make tables has an opportunity cost of $1,500. Choosing to convert the oak into tables means giving up the best alternatives that the firm could have purchased with the $1,500.

As a result, keep in mind that *all* of the resources that a firm uses—whether purchased from outside or already owned—have opportunity costs and thus economic costs. Economists refer to these two types of economic costs as *explicit costs* and *implicit costs*:

- A firm's **explicit costs** are the monetary payments it makes to those from whom it must purchase resources that it does not own. Because these costs involve an obvious cash transaction, they are referred to as explicit costs. Be sure to remember that explicit costs are opportunity costs because every monetary payment used to purchase outside resources necessarily involves forgoing the best alternatives that could have been purchased with the money.
- A firm's **implicit costs** are the opportunity costs of using the resources that it already owns to make the firm's own product rather than selling those resources to outsiders for cash. Because these costs are present but not obvious, they are referred to as implicit costs.

A firm's economic costs are the sum of its explicit costs and its implicit costs:

$$\text{Economic costs} = \text{explicit costs} + \text{implicit costs}$$

The following example makes clear how both explicit costs and implicit costs affect firm profits and firm behavior.

Accounting Profit and Normal Profit

Suppose that after many years working as a sales representative for a large T-shirt manufacturer, you decided to strike out on your own. After considering many potential business ventures, you decide to open a retail T-shirt shop. As we explain in Chapter 2, you will be providing two different economic resources to your new enterprise: labor and entrepreneurial ability. The part of your job that involves providing labor includes any of the routine tasks that are needed to help run the business—things like answering customer e-mails, taking inventory, and sweeping the floor. The part of your job that involves providing entrepreneurial ability includes any of the nonroutine tasks involved with organizing the business and directing its strategy—things like deciding on whether to use Internet ads or in-person events to promote your business, whether to include children's clothing in your product mix, and how to decorate your store to maximize its appeal to potential customers.

You begin providing entrepreneurial ability to your new firm by making some initial organizational decisions. You decide to work full time at your new business, so you quit your old job that paid you $22,000 per year. You invest $20,000 of savings that has been earning $1,000 per year. You decide that your new firm will occupy a small retail space that you own and had been previously renting out for $5,000 per year. Finally, you decide to hire one clerk to help you in the store. She agrees to work for you for $18,000 per year.

After a year in business, you total up your accounts and find the following:

Total sales revenue		$120,000
Cost of T-shirts	$40,000	
Clerk's salary	18,000	
Utilities	5,000	
Total (explicit) costs		63,000
Accounting profit		57,000

These numbers look very good. In particular, you are happy with your $57,000 **accounting profit,** the profit number that accountants calculate by subtracting total explicit costs from total sales revenue. This is the profit (or "net income") that would appear on your accounting statement and that you would report to the government for tax purposes.

But don't celebrate yet! Your $57,000 accounting profit overstates the economic success of your business because it ignores your implicit costs. Success is not defined as "having a total sales revenue that exceeds total explicit costs." Rather, the true measure of success is doing as well as you possibly can—that is, making more money in your new venture selling T-shirts than you could pursuing any other business venture.

To figure out whether you are achieving that goal, you must take into account all of your opportunity costs—both your implicit costs as well as your explicit costs. Doing so will indicate whether your new business venture is earning more money than what you could have earned in any other business venture.

To see how these calculations are made, let's continue with our example.

By providing your own financial capital, retail space, and labor, you incurred three different implicit costs during the year: $1,000 of forgone interest, $5,000 of forgone rent, and $22,000 of forgone wages. But don't forget that there is another implicit cost that you must also take account of—how much income you chose to forgo by applying your entrepreneurial abilities to your current retail T-shirt venture rather than applying them to other potential business ventures.

But what dollar value should we place on the size of the profits that you might have made if you had provided your entrepreneurial ability to one of those other ventures?

The answer is given by estimating a **normal profit,** the typical (or "normal") amount of accounting profit that you would most likely have earned in one of these other ventures. For the sake of argument, let us assume that with your particular set of skills and talents your entrepreneurial abilities would have on average yielded a normal profit of $5,000 in one of the other potential ventures. Knowing that value, we can take all of your implicit costs properly into account by subtracting them from your accounting profit:

Accounting profit		$57,000
Forgone interest	$ 1,000	
Forgone rent	5,000	
Forgone wages	22,000	
Forgone entrepreneurial income	5,000	
Total implicit costs		33,000
Economic profit		24,000

Economic Profit

After subtracting your $33,000 of implicit costs from your accounting profit of $57,000, we are left with an *economic profit* of $24,000.

Please distinguish clearly between accounting profit and economic profit. Accounting profit is the result of subtracting only explicit costs from revenue: *Accounting Profit = Revenue − Explicit Costs.* By contrast, **economic profit** is the result of subtracting all of your economic costs—both explicit costs and implicit costs—from revenue: *Economic Profit = Revenue − Explicit Costs − Implicit Costs.*

WORKED PROBLEMS

W9.1
Economic profit

By subtracting *all* of your economic costs from your revenue, you determine how your current business venture compares with your best alternative business venture. In our example, the fact that you are generating an economic profit of $24,000 means that you are making $24,000 more than you could expect to make in your best alternative business venture.

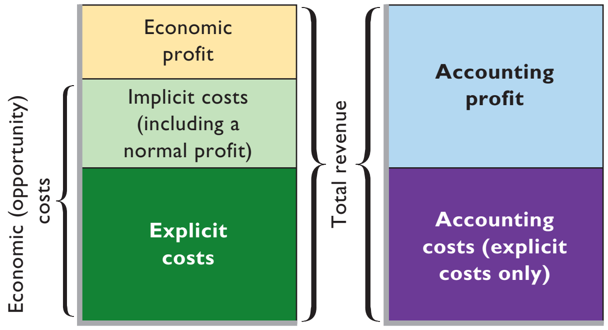
By contrast, suppose that you had instead done poorly in business, so that this year your firm generated an economic loss (a negative economic profit) of $8,000. This would mean that you were doing worse in your current venture than you could have done in your best alternative venture. You would, as a result, wish to switch to that alternative.

Generalizing this point, we see that there is an important behavioral threshold at $0 of economic profit. If a firm is breaking even (that is, earning exactly $0 of economic profit), then its entrepreneurs know that they are doing exactly as well as they could expect to do in their best alternative business venture. They are earning enough to cover all their explicit and implicit costs, including the normal profit that they could expect to earn in other business ventures. Thus, they have no incentive to change. By contrast, anyone running a positive economic profit knows they are doing better than they could in alternative ventures and will want to continue doing what they are doing or maybe even expand their business. And anyone running an economic loss (a negative economic profit) knows that they could do better by switching to something else.

It is for this reason that economists focus on economic profits rather than accounting profits. Simply put, economic profits direct how resources are allocated in the economy. Entrepreneurs running economic losses close their current businesses, thereby liberating the land, labor, capital, and entrepreneurial ability that they had been using. These resources are freed up to be used by firms that are generating positive economic profits or that are at least breaking even. Resources thus flow from producing goods and services with lower net benefits toward producing goods and services with higher net benefits. Allocative efficiency increases as firms are led by their profit signals to produce more of what consumers want the most.

Figure 9.1 shows the relationship among the various cost and profit concepts that we have just discussed. To test yourself, you might want to enter the cost numbers used in our example in the appropriate blocks.

FIGURE 9.1 Economic profit versus accounting profit. Economic profit is equal to total revenue less economic costs. Economic costs are the sum of explicit and implicit costs and include a normal profit to the entrepreneur. Accounting profit is equal to total revenue less accounting (explicit) costs.



Short Run and Long Run

When the demand for a firm's product changes, the firm's profitability may depend on how quickly it can adjust the amounts of the various resources it employs. It can easily and quickly adjust the quantities employed of many resources such as hourly labor, raw materials, fuel, and power. It needs much more time, however, to adjust its *plant capacity*—the size of the factory building, the amount of machinery and equipment, and other capital resources. In some heavy industries such as aircraft manufacturing, a firm may need several years to alter plant capacity. Because of these differences in adjustment time, economists find it useful to distinguish between two conceptual periods: the short run and the long run. We will discover that costs differ in these two time periods.

Short Run: Fixed Plant In microeconomics, the **short run** is a period too brief for a firm to alter its plant capacity, yet long enough to permit a change in the degree to which the plant's current capacity is used. The firm's plant capacity is fixed in the short run. However, the firm can vary its output by applying larger or smaller amounts of labor, materials, and other resources to that plant. It can use its existing plant capacity more or less intensively in the short run.

Long Run: Variable Plant In microeconomics, the **long run** is a period long enough for a firm to adjust the quantities of all the resources that it employs, including plant capacity. From the industry's viewpoint, the long run also includes enough time for existing firms to dissolve and leave the industry or for new firms to be created and enter the industry. While the short run is a "fixed-plant" period, the long run is a "variable-plant" period.

Illustrations If Boeing hires 100 extra workers for one of its commercial airline plants or adds an entire shift of workers, we are speaking of the short run. If it adds a new production facility and installs more equipment, we are referring to the long run. The first situation is a *short-run adjustment;* the second is a *long-run adjustment.*

The short run and the long run are conceptual periods rather than calendar time periods. In light-manufacturing industries, changes in plant capacity may be accomplished almost overnight. A small T-shirt manufacturer can increase its plant capacity in a matter of days by ordering and installing two or three new cutting tables and several extra sewing machines. But for heavy industry the long run is a different matter. Shell Oil may need several years to construct a new gasoline refinery.

QUICK REVIEW 9.1

- Explicit costs are money payments a firm makes to outside suppliers of resources; implicit costs are the opportunity costs associated with a firm's use of resources it owns.
- Normal profit is the implicit cost of entrepreneurship. Economic profit is total revenue less all explicit and implicit costs, including a normal profit.
- In the short run, a firm's plant capacity is fixed; in the long run, a firm can vary its plant size and firms can enter or leave the industry.

Short-Run Production Relationships

LO9.2 Relate the law of diminishing returns to a firm's short-run production costs.

A firm's costs of producing a specific output depend on both the prices and the quantities of the resources (inputs) needed to produce that output. Resource supply and demand determine resource prices. The technological aspects of production, specifically the relationships between inputs and output, determine the quantities of resources needed. Our focus will be on the *labor*-output relationship, given a fixed plant capacity. But before examining that relationship, we need to define three terms:

- **Total product (TP)** is the total quantity, or total output, of a particular good or service produced.
- **Marginal product (MP)** is the extra output or added product associated with adding a unit of a variable resource, in this case labor, to the production process. Thus,

$$\text{Marginal product} = \frac{\text{change in total product}}{\text{change in labor input}}$$

- **Average product (AP),** also called labor productivity, is output per unit of labor input:

$$\text{Average product} = \frac{\text{total product}}{\text{units of labor}}$$

In the short run, a firm can for a time increase its output by adding units of labor to its fixed plant. But by how much will output rise when it adds more labor? And why do we say "for a time"?

Law of Diminishing Returns

The answers are provided in general terms by the **law of diminishing returns.** This law assumes that technology is fixed and thus the techniques of production do not change. It states that as successive units of a variable resource (say, labor) are added to a fixed resource (say, capital or land), beyond some point the extra, or marginal, product that can be attributed to each additional unit of the variable resource will decline. For example, if additional workers are hired to work with a constant amount of capital equipment, output will eventually rise by smaller and smaller amounts as more workers are hired.

ORIGIN OF THE IDEA

O9.1
Law of diminishing returns

Rationale Suppose a farmer has a fixed resource—80 acres of land—planted in corn. If the farmer does not cultivate the cornfields (clear the weeds) at all, the yield will be 40 bushels per acre. If he cultivates the land once, output may rise to 50 bushels per acre. A second cultivation may increase output to 57 bushels per acre, a third to 61, and a fourth to 63. Succeeding cultivations will add less and less to the land's yield. If this were not so, the world's needs for corn could be fulfilled by extremely intense cultivation of this single 80-acre plot of land. Indeed, if diminishing returns did not occur, the world could be fed out of a flowerpot. Why not? Just keep adding more seed, fertilizer, and harvesters!

The law of diminishing returns also holds true in nonagricultural industries. Assume a wood shop is manufacturing furniture frames. It has a specific amount of equipment such as lathes, planes, saws, and sanders. If this shop hired just one or two workers, total output and productivity (output per worker) would be very low. The workers would have to perform many different jobs, and the advantages of specialization would not be realized. Time would be lost in switching from one job to another, and machines would stand idle much of the time. In short, the plant would be understaffed, and production would be inefficient because there would be too much capital relative to the amount of labor.

The shop could eliminate those difficulties by hiring more workers. Then the equipment would be more fully used, and workers could specialize on doing a single job. Time would no longer be lost switching from job to job. As more workers were added, production would become more efficient and the marginal product of each succeeding worker would rise.

But the rise could not go on indefinitely. Beyond a certain point, adding more workers would cause overcrowding. Since workers would then have to wait in line to use the machinery, they would be underused. Total output

CONSIDER THIS . . .

Diminishing Returns from Study

Here is a noneconomic example of a relationship between "inputs" and "output" that may help you better understand the idea of diminishing returns. Suppose for an individual that

Total course learning = *f* (intelligence, quality of course materials, instructor effectiveness, class time, and study time)

where *f* means "function of" or "depends on." So this relationship supposes that total course learning depends on intelligence (however defined), quality of course materials such as the textbook, the effectiveness of the instructor, the amount of class time, and the amount of personal study time outside the class.

For analytical purposes, let's assume that one's intelligence, the quality of course materials, the effectiveness of the instructor, and the amount of class time are *fixed*—meaning they do not change over the length of the course. Now let's add units of study time per day over the length of the course to "produce" greater course learning. The first hour of study time per day increases total course learning. Will the second hour enhance course learning by as much as the first? By how much will the third, fourth, fifth, . . . fifteenth hour of study per day contribute to total course learning relative to the *immediate previous hour*?

We think you will agree that eventually diminishing returns to course learning will set in as successive hours of study are added each day. At some point the marginal product of an extra hour of study time will decline and, at some further point, become zero.

This is also true of production relationships within firms. As successive units of a variable input (say, labor) are added to a fixed input (say, capital), the marginal product of the variable input eventually declines. In short, diminishing returns will occur sooner or later. Total product eventually will rise at a diminishing rate, reach a maximum, and then decline.

would increase at a diminishing rate because, given the fixed size of the plant, each worker would have less capital equipment to work with as more and more labor was hired. The marginal product of additional workers would decline because there would be more labor in proportion to the fixed amount of capital. Eventually, adding still more workers would cause so much congestion that marginal product would become negative and total product would decline. At the extreme, the addition of more and more labor would exhaust all the standing room, and total product would fall to zero.

Note that the law of diminishing returns assumes that all units of labor are of equal quality. Each successive worker is presumed to have the same innate ability, motor coordination, education, training, and work experience. Marginal product ultimately diminishes, but not because successive workers are less skilled or less energetic. It declines because the firm is using more workers relative to the amount of plant and equipment available.

Tabular Example Table 9.1 is a numerical illustration of the law of diminishing returns. Column 2 shows the total product, or total output, resulting from combining each level of a variable input (labor) in column 1 with a fixed amount of capital.

Column 3 shows the marginal product (MP), the change in total product associated with each additional unit of labor. Note that with no labor input, total product is zero; a plant with no workers will produce no output. The first three units of labor generate increasing marginal returns, with marginal products of 10, 15, and 20 units, respectively. But beginning with the fourth unit of labor, marginal product diminishes continuously, becoming zero with the seventh unit of labor and negative with the eighth.

WORKED PROBLEMS

W9.2

Total, marginal, and average product

Average product, or output per labor unit, is shown in column 4. It is calculated by dividing total product (column 2) by the number of labor units needed to produce it (column 1). At 5 units of labor, for example, AP is 14 (= 70/5).

Graphical Portrayal **Figure 9.2 (Key Graph)** shows the diminishing-returns data in Table 9.1 graphically and further clarifies the relationships between total, marginal, and average products. (Marginal product in Figure 9.2b is plotted halfway between the units of labor since it applies to the addition of each labor unit.)

Note first in Figure 9.2a that total product, TP, goes through three phases: It rises initially at an increasing rate; then it increases, but at a diminishing rate; finally, after reaching a maximum, it declines.

Geometrically, marginal product—shown by the MP curve in Figure 9.2b—is the slope of the total-product curve. Marginal product measures the change in total product associated with each succeeding unit of labor. Thus, the three phases of total product are also reflected

TABLE 9.1 **Total, Marginal, and Average Product: The Law of Diminishing Returns**

(1) Units of the Variable Resource (Labor)	(2) Total Product (TP)	(3) Marginal Product (MP), Change in (2)/Change in (1)		(4) Average Product (AP), (2)/(1)
0	0			—
1	10	10	Increasing marginal returns	10.00
2	25	15		12.50
3	45	20		15.00
4	60	15	Diminishing marginal returns	15.00
5	70	10		14.00
6	75	5		12.50
7	75	0	Negative marginal returns	10.71
8	70	−5		8.75

in marginal product. Where total product is increasing at an increasing rate, marginal product is rising. Here, extra units of labor are adding larger and larger amounts to total product. Similarly, where total product is increasing but at a decreasing rate, marginal product is positive but falling. Each additional unit of labor adds less to total product than did the previous unit. When total product is at a maximum, marginal product is zero. When total product declines, marginal product becomes negative.

Average product, AP (Figure 9.2b), displays the same tendencies as marginal product. It increases, reaches a maximum, and then decreases as more and more units of labor are added to the fixed plant. But note the relationship between marginal product and average product: Where marginal product exceeds average product, average product rises. And where marginal product is less than average product, average product declines. It follows that marginal product intersects average product where average product is at a maximum.

This relationship is a mathematical necessity. If you add a larger number to a total than the current average of that total, the average must rise. And if you add a smaller number to a total than the current average of that total, the average must fall. You raise your average examination grade only when your score on an additional (marginal) examination is greater than the average of all your past scores. You lower your average when your grade on an additional exam is below your current average. In our production example, when the amount an extra worker adds to total product exceeds the average product of all workers currently employed, average product will rise. Conversely, when the amount an extra worker adds to total product is less than the current average product, average product will decrease.

ORIGIN OF THE IDEA

O9.2

Production relationship

The law of diminishing returns is embodied in the shapes of all three curves. But, as our definition of the law of diminishing returns indicates, economists are most concerned with its effects on marginal product. The regions of increasing, diminishing, and negative marginal product (returns) are shown in Figure 9.2b.

Short-Run Production Costs

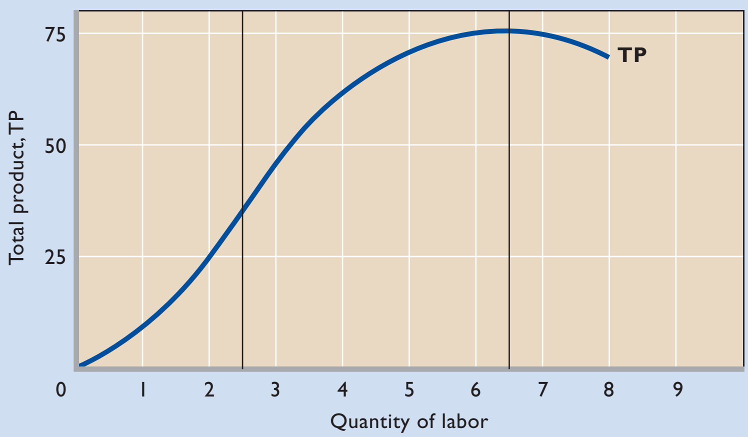
LO9.3 Describe the distinctions between fixed and variable costs and among total, average, and marginal costs.

Production information such as that provided in Table 9.1 and Figures 9.2a and 9.2b must be coupled with resource prices to determine the total and per-unit costs of producing various levels of output. We know that in the short run, resources associated with the firm's plant are fixed. Other resources, however, are variable in the short run. As a result, short-run costs can be either fixed or variable.

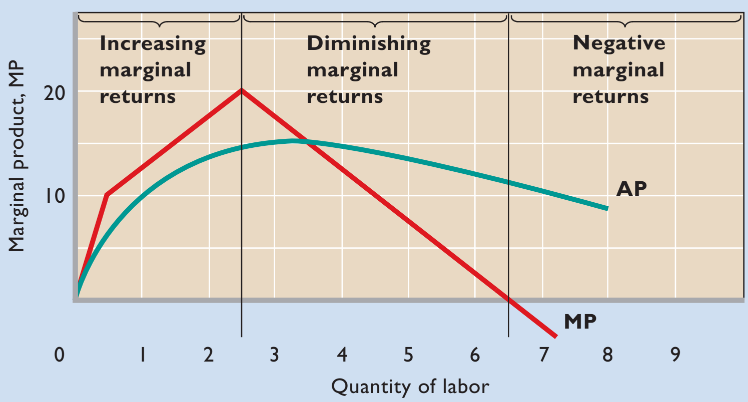
Fixed, Variable, and Total Costs

Let's see what distinguishes fixed costs, variable costs, and total costs from one another.

Fixed Costs **Fixed costs** are those costs that do not vary with changes in output. Fixed costs are associated with the very existence of a firm's plant and therefore must be paid even if its output is zero. Such costs as rental payments, interest on a firm's debts, a portion of depreciation on equipment and buildings, and insurance premiums are generally fixed costs; they are fixed and do not change even if a firm



(a)
Total product



(b)
Marginal and average product

FIGURE 9.2 The law of diminishing returns. (a) As a variable resource (labor) is added to fixed amounts of other resources (land or capital), the total product that results will eventually increase by diminishing amounts, reach a maximum, and then decline. (b) Marginal product is the change in total product associated with each new unit of labor. Average product is simply output per labor unit. Note that marginal product intersects average product at the maximum average product.

QUICK QUIZ FOR FIGURE 9.2

1. Which of the following is an assumption underlying these figures?
- **a.** Firms first hire "better" workers and then hire "poorer" workers.
- **b.** Capital and labor are both variable, but labor increases more rapidly than capital.
- **c.** Consumers will buy all the output (total product) produced.
- **d.** Workers are of equal quality.

2. Marginal product is:
- **a.** the change in total product divided by the change in the quantity of labor.
- **b.** total product divided by the quantity of labor.
- **c.** always positive.
- **d.** unrelated to total product.

3. Marginal product in graph (b) is zero when:
- **a.** average product in graph (b) stops rising.
- **b.** the slope of the marginal-product curve in graph (b) is zero.
- **c.** total product in graph (a) begins to rise at a diminishing rate.
- **d.** the slope of the total-product curve in graph (a) is zero.

4. Average product in graph (b):
- **a.** rises when it is less than marginal product.
- **b.** is the change in total product divided by the change in the quantity of labor.
- **c.** can never exceed marginal product.
- **d.** falls whenever total product in graph (a) rises at a diminishing rate.

Answers: 1. d; 2. a; 3. d; 4. a

TABLE 9.2 Total-, Average-, and Marginal-Cost Schedules for an Individual Firm in the Short Run

Total-Cost Data				Average-Cost Data			Marginal Cost
(1) Total Product (Q)	(2) Total Fixed Cost (TFC)	(3) Total Variable Cost (TVC)	(4) Total Cost (TC) TC = TFC + TVC	(5) Average Fixed Cost (AFC) $AFC = \frac{TFC}{Q}$	(6) Average Variable Cost (AVC) $AVC = \frac{TVC}{Q}$	(7) Average Total Cost (ATC) $ATC = \frac{TC}{Q}$	(8) Marginal Cost (MC) $MC = \frac{\text{change in TC}}{\text{change in } Q}$
0	$100	$ 0	$ 100				$ 90
1	100	90	190	$100.00	$90.00	$190.00	80
2	100	170	270	50.00	85.00	135.00	70
3	100	240	340	33.33	80.00	113.33	60
4	100	300	400	25.00	75.00	100.00	70
5	100	370	470	20.00	74.00	94.00	80
6	100	450	550	16.67	75.00	91.67	90
7	100	540	640	14.29	77.14	91.43	110
8	100	650	750	12.50	81.25	93.75	130
9	100	780	880	11.11	86.67	97.78	150
10	100	930	1,030	10.00	93.00	103.00	

produces more. In column 2 of Table 9.2 we assume that the firm's total fixed cost is $100. By definition, this fixed cost is incurred at all levels of output, including zero. The firm cannot avoid paying fixed costs in the short run.

Variable Costs **Variable costs** are those costs that change with the level of output. They include payments for materials, fuel, power, transportation services, most labor, and similar variable resources. In column 3 of Table 9.2 we find that the total of variable costs changes directly with output. But note that the increases in variable cost associated with succeeding one-unit increases in output are not equal. As production begins, variable cost will for a time increase by a decreasing amount; this is true through the fourth unit of output in Table 9.2. Beyond the fourth unit, however, variable cost rises by increasing amounts for succeeding units of output.

The reason lies in the shape of the marginal-product curve. At first, as in Figure 9.2b, marginal product is increasing, so smaller and smaller increases in the amounts of variable resources are needed to produce successive units of output. Hence the variable cost of successive units of output decreases. But when, as diminishing returns are encountered, marginal product begins to decline, larger and larger additional amounts of variable resources are needed to produce successive units of output. Total variable cost therefore increases by increasing amounts.

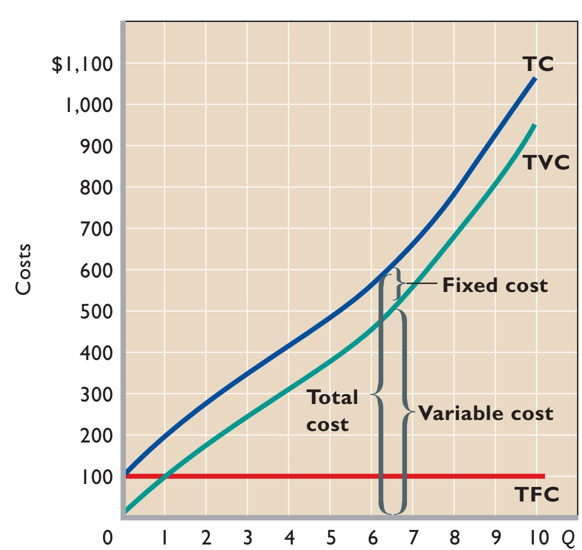
Total Cost **Total cost** is the sum of fixed cost and variable cost at each level of output:

$$TC = TFC + TVC$$

TC is shown in column 4 of Table 9.2. At zero units of output, total cost is equal to the firm's fixed cost. Then for each unit of the 10 units of production, total cost increases by the same amount as variable cost.

Figure 9.3 shows graphically the fixed-, variable-, and total-cost data given in Table 9.2. Observe that total variable cost, TVC, is measured vertically from the horizontal axis at each level of output. The amount of fixed

FIGURE 9.3 Total cost is the sum of fixed cost and variable cost. Total variable cost (TVC) changes with output. Total fixed cost (TFC) is independent of the level of output. The total cost (TC) at any output is the vertical sum of the fixed cost and variable cost at that output.



cost, shown as TFC, is added vertically to the total-variable-cost curve to obtain the points on the total-cost curve TC.

The distinction between fixed and variable costs is significant to the business manager. Variable costs can be controlled or altered in the short run by changing production levels. Fixed costs are beyond the business manager's current control; they are incurred in the short run and must be paid regardless of output level.

Per-Unit, or Average, Costs

Producers are certainly interested in their total costs, but they are equally concerned with per-unit, or average, costs. In particular, average-cost data are more meaningful for making comparisons with product price, which is always stated on a per-unit basis. Average fixed cost, average variable cost, and average total cost are shown in columns 5 to 7, Table 9.2.

AFC Average fixed cost (AFC) for any output level is found by dividing total fixed cost (TFC) by that amount of output (Q). That is,

$$\text{AFC} = \frac{\text{TFC}}{Q}$$

Because the total fixed cost is, by definition, the same regardless of output, AFC must decline as output increases. As output rises, the total fixed cost is spread over a larger and larger output. When output is just 1 unit in Table 9.2, TFC and AFC are the same at $100. But at 2 units of output, the total fixed cost of $100 becomes $50 of AFC or fixed cost per unit; then it becomes $33.33 per unit as $100 is spread over 3 units, and $25 per unit when spread over 4 units. This process is sometimes referred to as "spreading the overhead." Figure 9.4 shows that AFC graphs as a continuously declining curve as total output is increased.

AVC Average variable cost (AVC) for any output level is calculated by dividing total variable cost (TVC) by that amount of output (Q):

$$\text{AVC} = \frac{\text{TVC}}{Q}$$

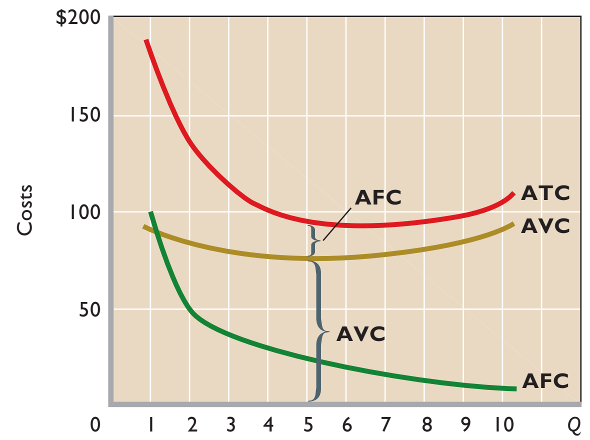
Due to increasing and then diminishing returns, AVC declines initially, reaches a minimum, and then increases again. A graph of AVC is a U-shaped or saucer-shaped curve, as shown in Figure 9.4.

Because total variable cost reflects the law of diminishing returns, so must AVC, which is derived from total variable cost. Because marginal returns increase initially, fewer and fewer additional variable resources are needed to produce each of the first four units of output. As a result, variable cost per unit declines. AVC hits a minimum with the fifth unit of output, and beyond that point AVC rises as diminishing returns require more and more variable resources to produce each additional unit of output.

Rephrased, production is relatively inefficient—and therefore costly—at low levels of output. Because the firm's fixed plant is understaffed, average variable cost is relatively high. As output expands, however, greater specialization and better use of the firm's capital equipment yield more efficiency, and variable cost per unit of output declines. As still more variable resources are added, a point is reached where crowding causes diminishing returns to set in. Once diminishing returns start, each additional unit of input does not increase output by as much as preceding units did. This means that AVC eventually increases.

You can verify the U or saucer shape of the AVC curve by returning to Table 9.1. Assume the price of labor is $10 per unit. Labor cost per unit of output is then $10 (the price per labor unit in this example) divided by average product (output per labor unit). Because we have assumed labor to be the only variable input, the labor cost per unit of output is the variable cost per unit of output, or AVC. When average product is initially low, AVC is high. As workers are added, average product rises and AVC falls. When average product is at its maximum, AVC is at its minimum. Then, as still more workers are added and average product declines, AVC rises. The "hump" of the

FIGURE 9.4 The average-cost curves. AFC falls as a given amount of fixed costs is apportioned over a larger and larger output. AVC initially falls because of increasing marginal returns but then rises because of diminishing marginal returns. Average total cost (ATC) is the vertical sum of average variable cost (AVC) and average fixed cost (AFC).



average-product curve is reflected in the saucer or U shape of the AVC curve. As you will soon see, the two are mirror images of each other.

ATC **Average total cost (ATC)** for any output level is found by dividing total cost (TC) by that output (Q) or by adding AFC and AVC at that output:

$$\text{ATC} = \frac{\text{TC}}{Q} = \frac{\text{TFC}}{Q} + \frac{\text{TVC}}{Q} = \text{AFC} + \text{AVC}$$

Graphically, ATC can be found by adding vertically the AFC and AVC curves, as in Figure 9.4. Thus the vertical distance between the ATC and AVC curves measures AFC at any level of output.

Marginal Cost

One final and very crucial cost concept remains: **Marginal cost (MC)** is the extra, or additional, cost of producing one more unit of output. MC can be determined for each added unit of output by noting the change in total cost entailed by that unit's production:

$$\text{MC} = \frac{\text{change in TC}}{\text{change in } Q}$$

Calculations In column 4 of Table 9.2, production of the first unit of output increases total cost from \$100 to \$190. Therefore, the additional, or marginal, cost of that first unit is \$90 (column 8). The marginal cost of the second unit is \$80 (= \$270 − \$190); the MC of the third is \$70 (= \$340 − \$270); and so forth. The MC for each of the 10 units of output is shown in column 8.

MC can also be calculated from the total-variable-cost column because the only difference between total cost and total variable cost is the constant amount of fixed costs (\$100). Thus, the change in total cost and the change in total variable cost associated with each additional unit of output are always the same.

Marginal Decisions Marginal costs are costs the firm can control directly and immediately. Specifically, MC designates all the cost incurred in producing the last unit of output. Thus, it also designates the cost that can be "saved" by not producing that last unit. Average-cost figures do not provide this information. For example, suppose the firm is undecided whether to produce 3 or 4 units of output. At 4 units Table 9.2 indicates that ATC is \$100. But the firm does not increase its total costs by \$100 by producing the fourth unit, nor does it save \$100 by not producing that unit. Rather, the change in costs involved here is only \$60, as the MC column in Table 9.2 reveals.

WORKED PROBLEMS

W9.3
Per-unit cost

A firm's decisions as to what output level to produce are typically marginal decisions, that is, decisions to produce a few more or a few less units. Marginal cost is the change in costs when one more or one less unit of output is produced. When coupled with marginal revenue (which, as you will see in Chapter 10, indicates the change in revenue from one more or one less unit of output), marginal cost allows a firm to determine if it is profitable to expand

CONSIDER THIS . . .

Ignoring Sunk Costs

It is a deep-seated human tendency to drag past costs—so called sunk costs—into marginal-benefit versus marginal-cost calculations. Doing so is known as the *sunk cost fallacy.*

As an example of this error, suppose a family that's on vacation stops at a roadside stand to buy some apples. After driving a bit, the family discovers that the apples are mushy and gross. Would it be logical for the father to insist that everyone eat the apples "because we paid a premium price for them"?

Absolutely not. In making a new decision, you should ignore all costs that are not affected by that new decision. The prior bad decision (in retrospect) to buy the apples should not dictate a subsequent decision for which marginal benefit is less than marginal cost.

Consider a business example. Suppose that a firm spends \$1 million on R&D to bring out a new product, only to discover that the product sells very poorly. Should the firm continue to produce the product at a loss even when there is no realistic hope for future success? Obviously, it should not. In making this decision, the firm should realize that the amount it spent developing the product is irrelevant; it should stop production and cut its losses.

The emotional tendency that drives the sunk cost fallacy is the desire to "get one's money's worth" out of a past expenditure. But giving in to that emotion can lead to "throwing good money after bad." Instead, you should ignore all past costs and focus solely on those that depend on the decision at hand.

KEY GRAPH

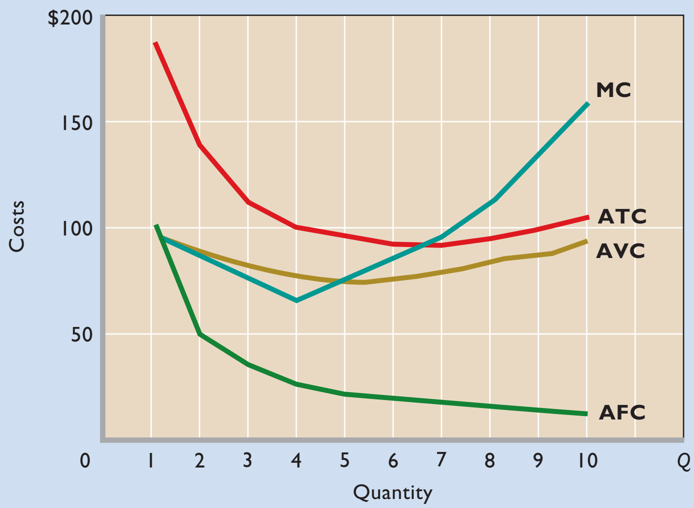


FIGURE 9.5 The relationship of the marginal-cost curve to the average-total-cost and average-variable-cost curves. The marginal-cost (MC) curve cuts through the average-total-cost (ATC) curve and the average-variable-cost (AVC) curve at their minimum points. When MC is below average total cost, ATC falls; when MC is above average total cost, ATC rises. Similarly, when MC is below average variable cost, AVC falls; when MC is above average variable cost, AVC rises.

QUICK QUIZ FOR FIGURE 9.5

1. The marginal-cost curve first declines and then increases because of:
- **a.** increasing, then diminishing, marginal utility.
- **b.** the decline in the gap between ATC and AVC as output expands.
- **c.** increasing, then diminishing, marginal returns.
- **d.** constant marginal revenue.

2. The vertical distance between ATC and AVC measures:
- **a.** marginal cost.
- **b.** total fixed cost.
- **c.** average fixed cost.
- **d.** economic profit per unit.

3. ATC is:
- **a.** AVC − AFC.
- **b.** MC + AVC.
- **c.** AFC + AVC.
- **d.** (AFC + AVC) + *Q*.

4. When the marginal-cost curve lies:
- **a.** above the ATC curve, ATC rises.
- **b.** above the AVC curve, ATC rises.
- **c.** below the AVC curve, total fixed cost increases.
- **d.** below the ATC curve, total fixed cost falls.

Answers: 1. c; 2. c; 3. c; 4. a

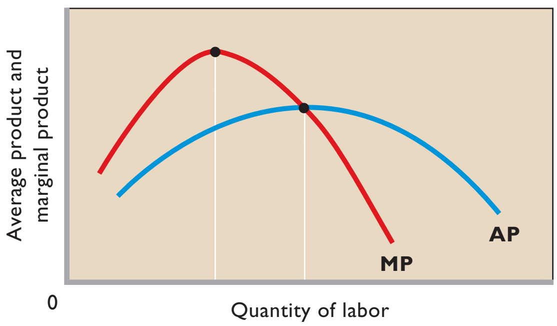
or contract its production. The analysis in the next four chapters focuses on those marginal calculations.

Graphical Portrayal Marginal cost is shown graphically in **Figure 9.5 (Key Graph).** Marginal cost at first declines sharply, reaches a minimum, and then rises rather abruptly. This reflects the fact that variable cost, and therefore total cost, increase at first by decreasing amounts and then by increasing amounts (see columns 3 and 4, Table 9.2).

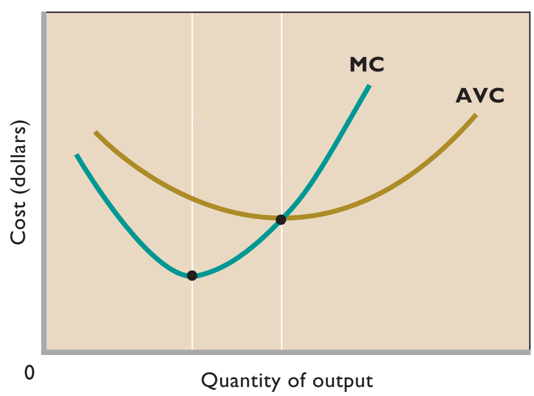
MC and Marginal Product The marginal-cost curve's shape is a consequence of the law of diminishing returns. Looking back at Table 9.1, we can see the relationship between marginal product and marginal cost. If all units of a variable resource (here labor) are hired at the same price, the marginal cost of each extra unit of output will fall as long as the marginal product of each additional worker is rising. This is true because marginal cost is the (constant) cost of an extra worker divided by his or her marginal product. Therefore, in Table 9.1, suppose that each worker can be hired for $10. Because the first worker's marginal product is 10 units of output, and hiring this worker increases the firm's costs by $10, the marginal cost of each of these 10 extra units of output is $1 (= $10/10 units). The second worker also increases costs by $10, but the marginal product is 15, so the marginal cost of each of these 15 extra units of output is $0.67 (= $10/15 units). Similarly, the MC of each of the 20 extra units of output contributed by the third worker is $.50 (= $10/20 units). To generalize, as long as marginal product is rising, marginal cost will fall.

But with the fourth worker diminishing returns set in and marginal cost begins to rise. For the fourth worker, marginal cost is $0.67 (= $10/15 units); for the fifth worker, MC is $1 ($10/10 units); for the sixth, MC is $2 (= $10/5 units); and so on. If the price (cost) of the variable resource

FIGURE 9.6 The relationship between productivity curves and cost curves. The marginal-cost (MC) curve and the average-variable-cost (AVC) curve in (b) are mirror images of the marginal-product (MP) and average-product (AP) curves in (a). Assuming that labor is the only variable input and that its price (the wage rate) is constant, then when MP is rising, MC is falling, and when MP is falling, MC is rising. Under the same assumptions, when AP is rising, AVC is falling, and when AP is falling, AVC is rising.



(a)
Production curves



(b)
Cost curves

remains constant, increasing marginal returns will be reflected in a declining marginal cost, and diminishing marginal returns in a rising marginal cost. The MC curve is a mirror reflection of the marginal-product curve. As you can see in Figure 9.6, when marginal product is rising, marginal cost is necessarily falling. When marginal product is at its maximum, marginal cost is at its minimum. And when marginal product is falling, marginal cost is rising.

Relation of MC to AVC and ATC Figure 9.5 shows that the marginal-cost curve MC intersects both the AVC and the ATC curves at their respective minimum points. As noted earlier, this marginal-average relationship is a mathematical necessity, which a simple illustration will reveal. Suppose an NBA basketball player has scored an average of 20 points a game over the first three games of the season. Now, whether his average rises or falls as a result of playing a fourth (marginal) game will depend on whether the additional points he scores in that game are fewer or more than his current 20-point average. If in the fourth game he scores fewer than 20 points, his average will fall. For example, if he scores 16 points in the fourth game, his total points will rise from 60 to 76 and his average will fall from 20 to 19 (= 76/4). Conversely, if in the fourth (marginal) game he scores more than 20 points, say, 24, his total will increase from 60 to 84 and his average will rise from 20 to 21 (= 84/4).

So it is with costs. When the amount (the marginal cost) added to total cost is less than the current average total cost, ATC will fall. Conversely, when the marginal cost exceeds ATC, ATC will rise. This means in Figure 9.5 that as long as MC lies below ATC, ATC will fall, and whenever MC lies above ATC, ATC will rise. Therefore, at the point of intersection where MC equals ATC, ATC has just ceased to fall but has not yet begun to rise. This, by definition, is the minimum point on the ATC curve. The marginal-cost curve intersects the average-total-cost curve at the ATC curve's minimum point.

Marginal cost can be defined as the addition either to total cost or to total variable cost resulting from one more unit of output; thus this same rationale explains why the MC curve also crosses the AVC curve at the AVC curve's minimum point. No such relationship exists between the MC curve and the average-fixed-cost curve because the two are not related; marginal cost includes only those costs that change with output, and fixed costs by definition are those that are independent of output.

Shifts of the Cost Curves

Changes in either resource prices or technology will cause costs to change and cost curves to shift. If fixed costs double from $100 to $200, the AFC curve in Figure 9.5 would be shifted upward. At each level of output, fixed costs are higher. The ATC curve would also move upward because AFC is a component of ATC. But the positions of the AVC and MC curves would be unaltered because their locations are based on the prices of variable rather than fixed resources. However, if the price (wage) of labor or some other variable input rose, AVC, ATC, and MC would rise and those cost curves would all shift upward. The AFC curve would remain in place because fixed costs have not changed. And, of course, reductions in the prices of fixed or variable resources would reduce costs and produce shifts of the cost curves exactly opposite to those just described.

The discovery of a more efficient technology would increase the productivity of all inputs. The cost figures in Table 9.2 would all be lower. To illustrate, if labor is the only variable input, if wages are $10 per hour, and if average product is 10 units, then AVC would be $1. But if a technological improvement increases the average product of labor to 20 units, then AVC will decline to $0.50. More generally, an upward shift in the productivity curves shown in Figure 9.6a means a downward shift in the cost curves portrayed in Figure 9.6b.

QUICK REVIEW 9.2

- The law of diminishing returns indicates that, beyond some point, output will increase by diminishing amounts as more units of a variable resource (labor) are added to a fixed resource (capital).
- In the short run, the total cost of any level of output is the sum of fixed and variable costs (TC = TFC + TVC).
- Average fixed, average variable, and average total costs are fixed, variable, and total costs per unit of output; marginal cost is the extra cost of producing one more unit of output.
- Average fixed cost declines continuously as output increases; the average-variable-cost and average-total-cost curves are U-shaped, reflecting increasing and then diminishing returns; the marginal-cost curve falls but then rises, intersecting both the average-variable-cost curve and the average-total-cost curve at their minimum points.

Long-Run Production Costs

LO9.4 Use economies of scale to link a firm's size and its average costs in the long run.

In the long run an industry and its individual firms can undertake all desired resource adjustments. That is, they can change the amount of all inputs used. The firm can alter its plant capacity; it can build a larger plant or revert to a smaller plant than that assumed in Table 9.2. The industry also can change its overall capacity; the long run allows sufficient time for new firms to enter or for existing firms to leave an industry. We will discuss the impact of the entry and exit of firms to and from an industry in the next chapter; here we are concerned only with changes in plant capacity made by a single firm. Let's couch our analysis in terms of average total cost (ATC), making no distinction between fixed and variable costs because all resources, and therefore all costs, are variable in the long run.

Firm Size and Costs

Suppose a manufacturer with a single plant begins on a small scale and, as the result of successful operations, expands to successively larger plant sizes with larger output capacities. What happens to average total cost as this occurs? For a time, successively larger plants will reduce average total cost. However, eventually the building of a still larger plant will cause ATC to rise.

Figure 9.7 illustrates this situation for five possible plant sizes. ATC-1 is the short-run average-total-cost curve for the smallest of the five plants, and ATC-5, the curve for the largest. Constructing larger plants will lower the minimum average total costs through plant size 3. But then larger plants will mean higher minimum average total costs.

The Long-Run Cost Curve

The vertical lines perpendicular to the output axis in Figure 9.7 indicate the outputs at which the firm should change plant size to realize the lowest attainable average total costs of production. These are the outputs at which the per-unit costs for a larger plant drop below those for the current, smaller plant. For all outputs up to 20 units, the lowest average total costs are attainable with plant size 1. However, if the firm's volume of sales expands beyond

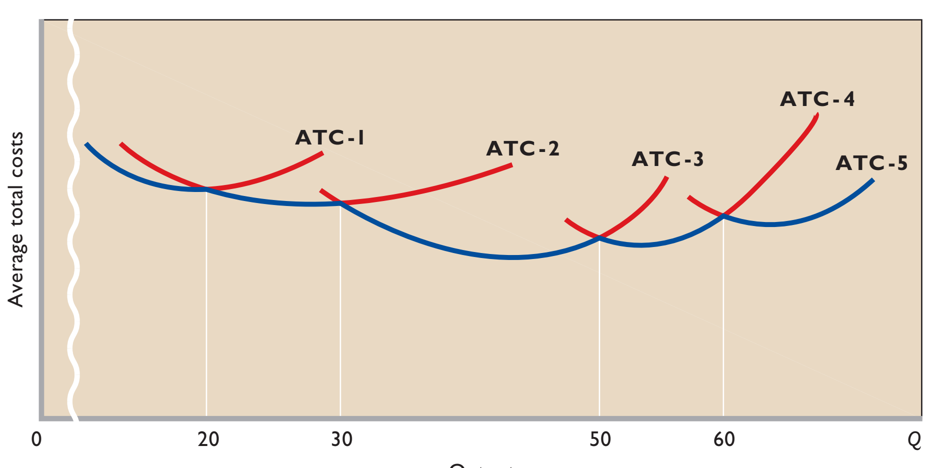


FIGURE 9.7 The long-run average-total-cost curve: five possible plant sizes. The long-run average-total-cost curve is made up of segments of the short-run cost curves (ATC-1, ATC-2, etc.) of the various-size plants from which the firm might choose. Each point on the bumpy planning curve shows the lowest unit cost attainable for any output when the firm has had time to make all desired changes in its plant size.

KEY GRAPH

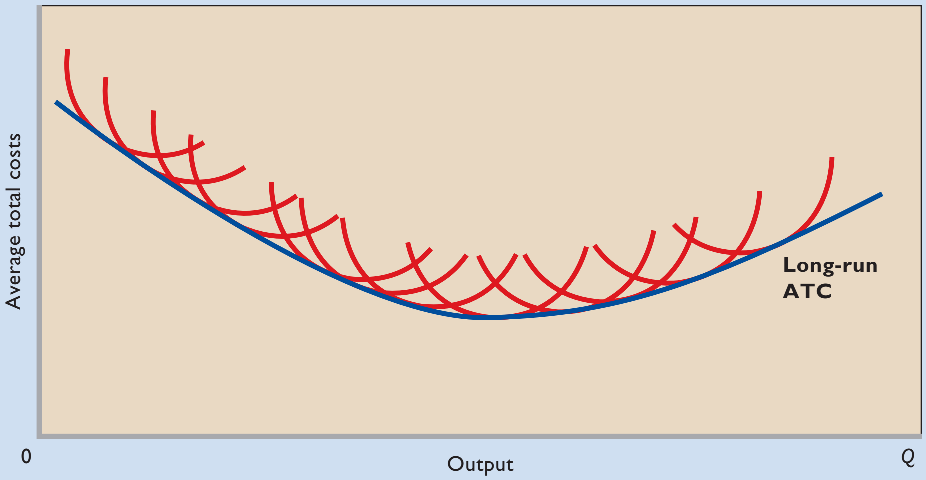


FIGURE 9.8 The long-run average-total-cost curve: unlimited number of plant sizes. If the number of possible plant sizes is very large, the long-run average-total-cost curve approximates a smooth curve. Economies of scale, followed by diseconomies of scale, cause the curve to be U-shaped.

QUICK QUIZ FOR FIGURE 9.8

1. The unlabeled red curves in this figure illustrate the:
- **a.** long-run average-total-cost curves of various firms constituting the industry.
- **b.** short-run average-total-cost curves of various firms constituting the industry.
- **c.** short-run average-total-cost curves of various plant sizes available to a particular firm.
- **d.** short-run marginal-cost curves of various plant sizes available to a particular firm.

2. The unlabeled red curves in this figure derive their shapes from:
- **a.** decreasing, then increasing, short-run returns.
- **b.** increasing, then decreasing, short-run returns.
- **c.** economies, then diseconomies, of scale.
- **d.** diseconomies, then economies, of scale.

3. The long-run ATC curve in this figure derives its shape from:
- **a.** decreasing, then increasing, short-run returns.
- **b.** increasing, then decreasing, short-run returns.
- **c.** economies, then diseconomies, of scale.
- **d.** diseconomies, then economies, of scale.

4. The long-run ATC curve is often called the firm's:
- **a.** planning curve.
- **b.** capital-expansion path.
- **c.** total-product curve.
- **d.** production possibilities curve.

Answers: 1. c; 2. b; 3. c; 4. a

20 units but less than 30, it can achieve lower per-unit costs by constructing a larger plant, size 2. Although total cost will be higher at the expanded levels of production, the cost per unit of output will be less. For any output between 30 and 50 units, plant size 3 will yield the lowest average total costs. From 50 to 60 units of output, the firm must build the size-4 plant to achieve the lowest unit costs. Lowest average total costs for any output over 60 units require construction of the still larger plant, size 5.

Tracing these adjustments, we find that the long-run ATC curve for the enterprise is made up of segments of the short-run ATC curves for the various plant sizes that can be constructed. The long-run ATC curve shows the lowest average total cost at which *any output level* can be produced after the firm has had time to make all appropriate adjustments in its plant size. In Figure 9.7 the blue, bumpy curve is the firm's long-run ATC curve or, as it is often called, the firm's *planning curve*.

In most lines of production the choice of plant size is much wider than in our illustration. In many industries the number of possible plant sizes is virtually unlimited, and in time quite small changes in the volume of output will lead to changes in plant size. Graphically, this implies an unlimited number of short-run ATC curves, one for each output level, as suggested by **Figure 9.8 (Key Graph).** Then, rather than being made up of segments of short-run ATC curves as in Figure 9.7, the long-run ATC curve is made up of all the points of tangency of the unlimited number of short-run ATC curves from which the long-run ATC curve is derived. Therefore, the planning curve is smooth rather than bumpy. Each point on it tells us the minimum ATC of producing the corresponding level of output.

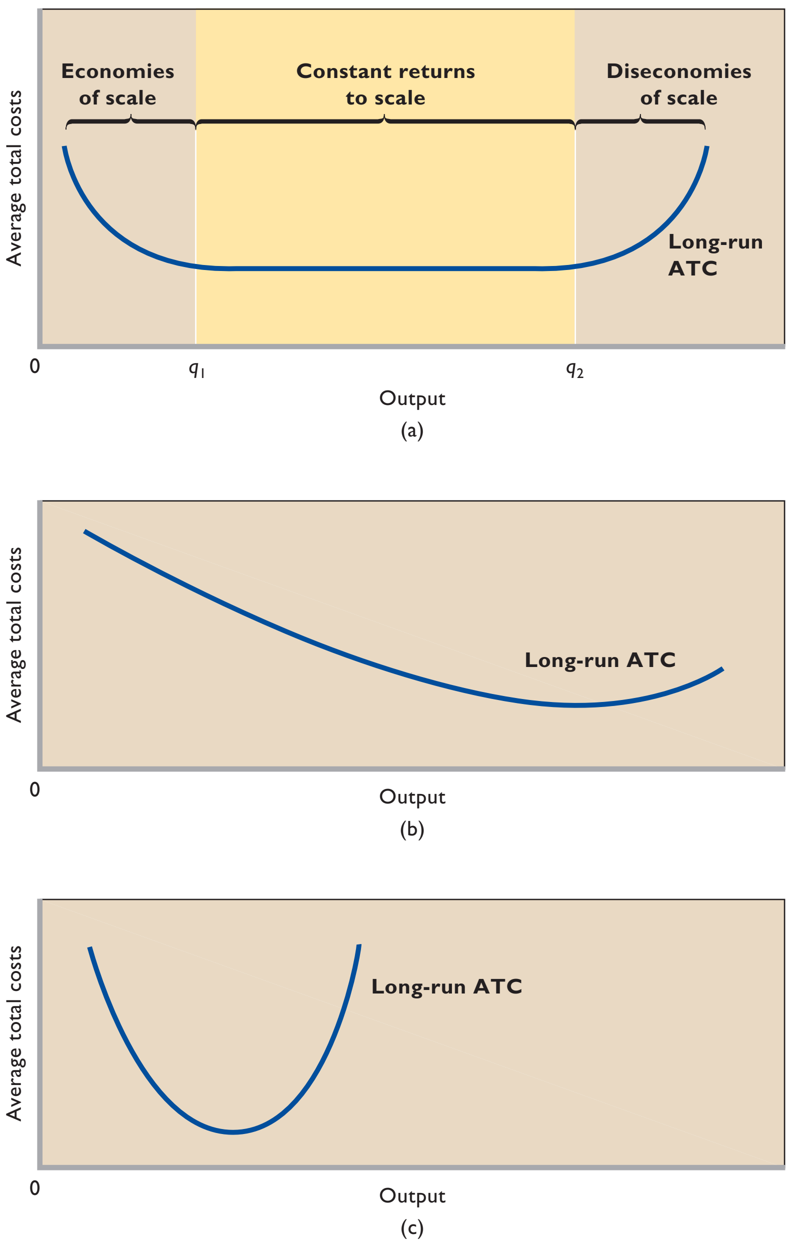


FIGURE 9.9 Various possible long-run average-total-cost curves. (a) Economies of scale are rather rapidly obtained as plant size rises, and diseconomies of scale are not encountered until a considerably large scale of output has been achieved. Thus, long-run average total cost is constant over a wide range of output. (b) Economies of scale are extensive, and diseconomies of scale occur only at very large outputs. Average total cost therefore declines over a broad range of output. (c) Economies of scale are exhausted quickly, followed immediately by diseconomies of scale. Minimum ATC thus occurs at a relatively low output.

Economies and Diseconomies of Scale

We have assumed that, for a time, larger and larger plant sizes will lead to lower unit costs but that, beyond some point, successively larger plants will mean higher average total costs. That is, we have assumed the long-run ATC curve is U-shaped. But why should this be? It turns out that the U shape is caused by economies and diseconomies of large-scale production, as we explain in a moment. But before we do, please understand that the U shape of the long-run average-total-cost curve *cannot* be the result of rising resource prices or the law of diminishing returns. First, our discussion assumes that resource prices are constant. Second, the law of diminishing returns does not apply to production in the long run. This is true because the law of diminishing returns only deals with situations in which a productive resource or input is held constant. Under our definition of "long run," all resources and inputs are variable.

Economies of Scale **Economies of scale,** or economies of mass production, explain the downsloping part of the long-run ATC curve, as indicated in Figure 9.9, graphs (a), (b), and (c). As plant size increases, a number of factors will for a time lead to lower average costs of production.

Labor Specialization Increased specialization in the use of labor becomes more achievable as a plant increases in size. Hiring more workers means jobs can be divided and subdivided. Each worker may now have just one task to perform instead of five or six. Workers can work full-time on the tasks for which they have special skills. By contrast, skilled machinists in a small plant may spend half their time performing unskilled tasks, leading to higher production costs.

Further, by working at fewer tasks, workers become even more proficient at those tasks. The jack-of-all-trades doing five or six jobs is not likely to be efficient in any of them. Concentrating on one task, the same worker may become highly efficient.

Finally, greater labor specialization eliminates the loss of time that occurs whenever a worker shifts from one task to another.

Managerial Specialization Large-scale production also means better use of, and greater specialization in, management. A supervisor who can handle 20 workers is underused in a small plant that employs only 10 people. The production staff could be doubled with no increase in supervisory costs.

Small firms cannot use management specialists to best advantage. For example, a marketing specialist working in a small plant may have to spend some of her time on functions outside of her area of expertise—for example, accounting, personnel, and finance. A larger scale of operations would allow her to supervise marketing full time, while other specialists perform other managerial functions. Greater productivity and efficiency, along with lower unit costs, would be the net result.

Efficient Capital Small firms often cannot afford the most efficient equipment. In many lines of production such machinery is available only in very large and extremely expensive units. Furthermore, effective use of the equipment demands a high volume of production, and that again requires large-scale producers.

In the automobile industry the most efficient fabrication method employs robotics and elaborate assembly-line equipment. Effective use of this equipment demands an annual output of several hundred thousand automobiles. Only very large-scale producers can afford to purchase and use this equipment efficiently. The small-scale producer is faced with a dilemma. To fabricate automobiles using other equipment is inefficient and therefore more costly per unit. But so, too, is buying and underutilizing the equipment used by the large manufacturers. Because it cannot spread the high equipment cost over very many units of output, the small-scale producer will be stuck with high costs per unit of output.

Other Factors Many products entail design and development costs, as well as other "start-up" costs, which must be incurred regardless of projected sales. These costs decline per unit as output is increased. Similarly, advertising costs decline per auto, per computer, per stereo system, and per box of detergent as more units are produced and sold. Also, the firm's production and marketing expertise usually rises as it produces and sells more output. This *learning by doing* is a further source of economies of scale.

All these factors contribute to lower average total costs for the firm that is able to expand its scale of operations. Where economies of scale are possible, an increase in all resources of, say, 10 percent will cause a more-than-proportionate increase in output of, say, 20 percent. The result will be a decline in ATC.

In many U.S. manufacturing industries, economies of scale have been of great significance. Firms that have expanded their scale of operations to obtain economies of mass production have survived and flourished. Those unable to expand have become relatively high-cost producers, doomed to struggle to survive.

Diseconomies of Scale In time the expansion of a firm may lead to diseconomies and therefore higher average total costs.

The main factor causing **diseconomies of scale** is the difficulty of efficiently controlling and coordinating a firm's operations as it becomes a large-scale producer. In a small plant a single key executive may make all the basic decisions for the plant's operation. Because of the firm's small size, the executive is close to the production line, understands the firm's operations, and can make efficient decisions because the small plant size requires only a relatively small amount of information to be examined and understood in order to optimize production.

This neat picture changes as a firm grows. One person cannot assemble, digest, and understand all the information essential to decision-making on a large scale. Authority must be delegated to many vice presidents, second vice presidents, and so forth. This expansion of the management hierarchy leads to problems of communication and cooperation, bureaucratic red tape, and the possibility that decisions will not be coordinated. At the same time, each new manager must be paid a salary. Thus, declining efficiency in making and executing decisions goes hand-in-hand with rising average total costs as bureaucracy expands beyond a certain point.

Also, in massive production facilities workers may feel alienated from their employers and care little about working efficiently. Opportunities to shirk, by avoiding work in favor of on-the-job leisure, may be greater in large plants than in small ones. Countering worker alienation and shirking may require additional worker supervision, which increases costs.

Where diseconomies of scale are operative, an increase in all inputs of, say, 10 percent will cause a less-than-proportionate increase in output of, say, 5 percent. As a consequence, ATC will increase. The rising portion of the long-run cost curves in Figure 9.9 illustrates diseconomies of scale.

Constant Returns to Scale In some industries a rather wide range of output may exist between the output at which economies of scale end and the output at which diseconomies of scale begin. That is, there may be a range of **constant returns to scale** over which long-run average cost does not change. The q_1q_2 output range of Figure 9.9a is an example. Here a given percentage increase in all inputs of, say, 10 percent will cause a proportionate 10 percent increase in output. Thus, in this range ATC is constant.

Minimum Efficient Scale and Industry Structure

Economies and diseconomies of scale are an important determinant of an industry's structure. Here we introduce the concept of **minimum efficient scale (MES),** which is the lowest level of output at which a firm can minimize long-run average costs. In Figure 9.9a that level occurs at q_1 units of output. Because of the extended range of constant returns to scale, firms producing substantially greater outputs could also realize the minimum attainable long-run average costs. Specifically, firms within the q_1 to q_2 range would be equally efficient. So we would not be surprised to find an industry with such cost conditions to be populated by firms of quite different sizes. The apparel, food processing, furniture, wood products, snowboard, banking, and small-appliance industries are examples. With an extended range of constant returns to scale, relatively large and relatively small firms can coexist in an industry and be equally successful.

Compare this with Figure 9.9b, where economies of scale continue over a wide range of outputs and diseconomies of scale appear only at very high levels of output. This pattern of declining long-run average total cost occurs in the automobile, aluminum, steel, and other heavy industries. The same pattern holds in several of the new industries related to information technology, for example, computer microchips, operating system software, and Internet service provision.

Given consumer demand, efficient production will be achieved with a few large-scale producers. Small firms cannot realize the minimum efficient scale and will not be able to compete. In the extreme, economies of scale might extend beyond the market's size, resulting in what is termed **natural monopoly,** a relatively rare market situation in which average total cost is minimized when only one firm produces the particular good or service.

Where economies of scale are few and diseconomies come into play quickly, the minimum efficient size occurs at a low level of output, as shown in Figure 9.9c. In such industries a particular level of consumer demand will support a large number of relatively small producers. Many retail trades and some types of farming fall into this category. So do certain kinds of light manufacturing such as the baking, clothing, and shoe industries. Fairly small firms are more efficient than larger-scale producers in such industries.

Our point here is that the shape of the long-run average-total-cost curve is determined by technology and the economies and diseconomies of scale that result. The shape of the long-run ATC curve, in turn, can be significant in determining whether an industry is populated by a relatively large number of small firms or is dominated by a few large producers, or lies somewhere in between.

ORIGIN OF THE IDEA

O9.3
Minimum efficient scale and natural monopoly

But we must be cautious in our assessment because industry structure does not depend on cost conditions alone. Government policies, the geographic size of markets, managerial strategy and skill, and other factors must be considered in explaining the structure of a particular industry.

QUICK REVIEW 9.3

- Most firms have U-shaped long-run average-total-cost curves, reflecting economies and then diseconomies of scale.
- Economies of scale are the consequence of greater specialization of labor and management, more efficient capital equipment, and the spreading of start-up costs over more units of output.
- Diseconomies of scale are caused by the problems of coordination and communication that arise in large firms.
- Minimum efficient scale (MES) is the lowest level of output at which a firm's long-run average total cost is at a minimum.

Applications and Illustrations

LO9.5 Give business examples of short-run costs, economies of scale, and minimum efficient scale (MES).

The business world offers many examples relating to short-run costs, economies of scale, and minimum efficient scale (MES). Here are just a few.

LAST WORD

3-D Printers

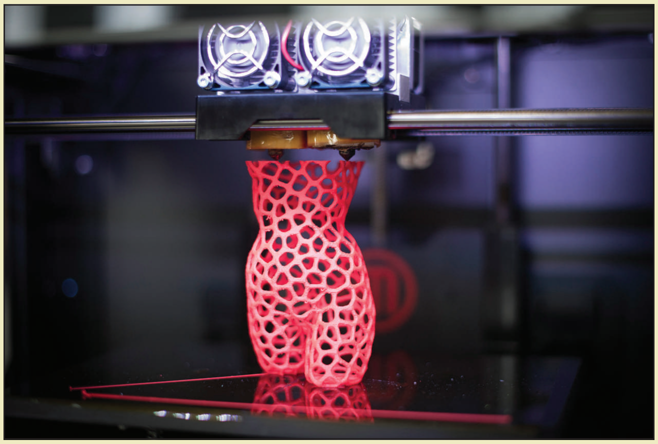
3-D Printers Are Poised to Replace Mass Production with Mass Customization.

Both a billionaire and your Average Joe can buy a pocketknife for $10. They can also both buy an iPhone for $199. And they can both purchase a new compact car for under $15,000.

The fact that all of these items are affordable to both a billionaire and your Average Joe is due to mass production and economies of scale. The iPhone, for instance, is one of the most complicated devices ever made. It contains cutting-edge technologies for graphics, voice recognition, battery length, screen durability, and many other features. Most of those technologies took hundreds of millions—if not billions—of dollars to develop and the factories that manufacture the iPhone and its components themselves cost many billions of dollars to set up. Yet, the iPhone is so inexpensive that Average Joes can afford to buy one.

That mass affordability is the result of mass production coupled with mass sales. Marginal costs are typically quite low with mass production. So if manufacturers can tap mass markets and sell their products in large numbers, they can achieve low per-unit costs by spreading the massive fixed costs (for developing the new technologies and setting up the factories) over many units. Doing so results in economies of scale, low average total costs per unit, and low prices that even average folks can afford.

Mass production and mass sales first became possible during the Industrial Revolution, which began in England during the late 1700s



and then spread through most of the rest of the world during the next two centuries. The Industrial Revolution occurred when steam-powered engines became powerful enough to drive factory equipment, propel ships, and pull trains. Engineers and inventors used steam power to automate factories and initiate the low-cost mass production of consumer goods. That process only accelerated when, in the late 19th century, the so-called Second Industrial Revolution saw electricity harnessed to drive factories and provide lighting.

Rising Gasoline Prices

As we discuss in the appendix to Chapter 3, changes in supply and demand often lead to rapid increases in the price of gasoline. Because gasoline is used to power the vast majority of all motor vehicles, including those used by businesses, increases in the price of gasoline lead to increases in firms' short-run variable costs, marginal costs, and average total costs. In terms of our analysis, their AVC, MC, and ATC curves all shift upward when an increase in the price of gasoline increases production costs.

The extent of these upward shifts depends upon the relative importance of gasoline as a variable input in the various firms' individual production processes. Package-delivery companies like FedEx that use a lot of gasoline-powered vehicles will see substantial upward shifts while software companies like Symantec (Norton) that mainly deliver their products through Internet downloads may see only small upward shifts.

Successful Start-Up Firms

The U.S. economy has greatly benefited over the past several decades from the explosive growth of scores of highly successful start-up firms. These firms typically reduce their costs by moving from higher to lower points on their short-run cost curves and by downward and to-the-right shifts of their short-run cost curves via economies of scale. That has certainly been the case for such former start-up firms as Intel (microchips), Starbucks (coffee), Microsoft (software), Dell (personal computers), Google (Internet searches), and Cisco Systems (Internet switching).

A major source of lower average total cost for rapidly growing firms is the ability to spread huge product development and advertising costs over a larger number of units of output. These firms also achieve economies of scale from learning by doing and through increased specialization of labor, management, and equipment. After

Mass sales, however, are not easy. They require large distribution networks, massive advertising budgets, and, perhaps most importantly, cheap ways of shipping products from factories to consumers. Thus, it was crucially important that transportation was also vastly improved during the first and second Industrial Revolutions. If not for better ships, smoother roads, and cheap transportation by railroad, transportation costs would have been so high that consumers would not have been able to afford mass-produced products shipped from distant factories.

Now, a new technology promises to deliver a Third Industrial Revolution that will feature not only low production costs but also zero transportation costs. Even better, both of those highly attractive features will be possible *even if you make only a single unit of a product*. In addition, each unit can be fully customized to a consumer's wants and needs. As a result, our world of affordable mass production may soon be replaced by a world of affordable mass customization.

The new technology is called additive manufacturing and it creates objects using computer-controlled devices known as "3-D printers." The 3-D (three-dimensional) printers contain a fine powder of metal or plastic particles that sit in a bin. A laser moves rapidly over the powder, the heat of its beam fusing small clumps of the powder together. Guided by a computerized blueprint, the rapidly moving laser can fuse a single layer of a complicated object together in just a few seconds.

The bin is then lowered a bit, another layer of powder is placed on top, and the laser again begins to shoot, this time fusing together both the previous layer as well as the current layer. Doing this over and over, one layer atop another, results in a solid object whose shape is limited only by the complexity of the blueprint. Any of the powder that is not struck by the laser and incorporated into the object is simply recycled for later use.

Because 3-D printers are inexpensive, they could potentially be located anywhere. Thus, there is no need to worry about transportation costs since objects could be manufactured by consumers in their own homes or in local workshops located only a short drive away. And because the powders are cheap and the machines only require a modest amount of electricity, anything that could be made using a 3-D printer would be inexpensive even if you were only making a single unit.

The First Industrial Revolution delivered low prices by spreading massive fixed costs over many units. The Third Industrial Revolution is set to deliver even lower prices by eliminating two types of costs—the massive fixed costs necessary to set up large factories and the transportation costs needed to ship resources to factories and then finished goods to consumers.

One major cost might still remain, however. That is the cost of paying people to make the blueprints that drive the 3-D printers. But just as digital file sharing has pushed the price of recorded music toward zero, many analysts suspect that digital file sharing will also drive the price of blueprints very low. If so, the cost of manufactured goods may soon plunge to levels even lower than what has been achieved through mass production.

So far, only relatively simple objects can be made with 3-D printers. But some engineers see a day in the not-so-distant future when it may be possible to create even complicated devices like an iPhone using additive manufacturing. If so, people will simply download inexpensive blueprints, make a few changes to customize the product, and then "print" what they want.

starting up, such firms experience declining average total costs over the years or even decades it takes them to eventually reach their respective MESs.

The Verson Stamping Machine

In 1996 Verson (a U.S. firm located in Chicago) introduced a 49-foot-tall metal-stamping machine that is the size of a house and weighs as much as 12 locomotives. This $30 million machine, which cuts and sculpts raw sheets of steel into automobile hoods and fenders, enables automakers to make new parts in just 5 minutes, compared with 8 hours for older stamping presses. A single machine is designed to make 5 million auto parts per year. So, to achieve the cost saving from the machine, an auto manufacturer must have sufficient auto production to use all these parts. By allowing the use of this cost-saving piece of equipment, large firm size achieves economies of scale.

The Daily Newspaper

Daily newspapers have been going bankrupt in rapid succession over the past several years as both advertising dollars and news readership have shifted to the Internet. The falling circulation numbers have caused average fixed costs to rise significantly as newspapers are forced to spread their substantial fixed costs over fewer and fewer papers. The spike in average fixed costs has, in turn, forced newspapers to sharply increase their prices. Between July 2007 and July 2009, for instance, the *New York Times* had to raise its cover price three times as advertising revenues plunged and fixed costs had to be spread over fewer and fewer papers. Starting at $1 per copy, the cover price had to be raised to $1.25, then $1.50, and then $2.00.

With readership continuing to fall, newspapers face an average-fixed-cost death spiral. The more they raise their prices, the less they will sell. But the less their sales, the higher their average fixed costs and thus the more they must raise their prices. As a result, printed newspapers could ultimately be a thing of the past, with both advertising and news delivery shifting mainly to the Internet.

Aircraft and Concrete Plants

Why are there only two plants in the United States (both operated by Boeing) that produce large commercial aircraft and thousands of plants (owned by hundreds of firms) that produce ready-mixed concrete? The simple answer is that MES is radically different in the two industries. Why is that? First, while economies of scale are extensive in assembling large commercial aircraft, they are only very modest in mixing concrete. Manufacturing airplanes is a complex process that requires huge facilities, thousands of workers, and very expensive, specialized machinery. Economies of scale extend to huge plant sizes. But mixing portland cement, sand, gravel, and water to produce concrete requires only a handful of workers and relatively inexpensive equipment. Economies of scale are exhausted at relatively small size.

The differing MESs also derive from the vastly different sizes of the geographic markets. The market for commercial airplanes is global, and aircraft manufacturers can deliver new airplanes anywhere in the world by flying them there. In contrast, the geographic market for a concrete plant is roughly the 50-mile radius within which the concrete can be delivered before it "sets up." So thousands of small concrete plants locate close to their customers in hundreds of small and large cities in the United States.

SUMMARY

LO9.1 Explain why economic costs include both explicit (revealed and expressed) costs and implicit (present but not obvious) costs.

The economic cost of using a resource to produce a good or service is the value or worth that the resource would have had in its best alternative use. Economic costs include explicit costs, which flow to resources owned and supplied by others, and implicit costs, which are payments for the use of self-owned and self-employed resources. One implicit cost is a normal profit to the entrepreneur. Economic profit occurs when total revenue exceeds total cost (= explicit costs + implicit costs, including a normal profit).

In the short run a firm's plant capacity is fixed. The firm can use its plant more or less intensively by adding or subtracting units of variable resources, but it does not have sufficient time in the short run to alter plant size.

LO9.2 Relate the law of diminishing returns to a firm's short-run production costs.

The law of diminishing returns describes what happens to output as a fixed plant is used more intensively. As successive units of a variable resource such as labor are added to a fixed plant, beyond some point the marginal product associated with each additional unit of a resource declines.

LO9.3 Describe the distinctions between fixed and variable costs and among total, average, and marginal costs.

Because some resources are variable and others are fixed, costs can be classified as variable or fixed in the short run. Fixed costs are independent of the level of output; variable costs vary with output. The total cost of any output is the sum of fixed and variable costs at that output.

Average fixed, average variable, and average total costs are fixed, variable, and total costs per unit of output. Average fixed cost declines continuously as output increases because a fixed sum is being spread over a larger and larger number of units of production. A graph of average variable cost is U-shaped, reflecting increasing returns followed by diminishing returns. Average total cost is the sum of average fixed and average variable costs; its graph is also U-shaped.

Marginal cost is the extra, or additional, cost of producing one more unit of output. It is the amount by which total cost and total variable cost change when one more or one less unit of output is produced. Graphically, the marginal-cost curve intersects the ATC and AVC curves at their minimum points.

Lower resource prices shift cost curves downward, as does technological progress. Higher input prices shift cost curves upward.

LO9.4 Use economies of scale to link a firm's size and its average costs in the long run.

The long run is a period of time sufficiently long for a firm to vary the amounts of all resources used, including plant size. In the long run all costs are variable. The long-run ATC, or planning, curve is composed of segments of short-run ATC curves, and it represents the various plant sizes a firm can construct in the long run.

The long-run ATC curve is generally U-shaped. Economies of scale are first encountered as a small firm expands. Greater specialization in the use of labor and management, the ability to use the most efficient equipment, and the spreading of start-up costs among more units of output all contribute to economies of scale. As the firm continues to grow, it will encounter diseconomies of scale stemming from the managerial complexities that accompany large-scale production. The output ranges over which economies and diseconomies of scale occur in an industry are often an important determinant of the structure of that industry.

A firm's minimum efficient scale (MES) is the lowest level of output at which it can minimize its long-run average cost. In some industries, MES occurs at such low levels of output that numerous firms can populate the industry. In other industries, MES occurs at such high output levels that only a few firms can exist in the long run.

LO9.5 Give business examples of short-run costs, economies of scale, and minimum efficient scale (MES).

Rising gasoline prices increase (shift upward) the AVC, ATC, and MC cost curves of firms like FedEx that use gasoline as an input in their production processes.

Starbucks, Facebook, and many other successful start-up firms that experienced rapid growth reduced costs and shifted their cost curves down and to the right by spreading product-development and advertising costs over larger numbers of units and by exploiting the economies of scale that can be generated through learning by doing and increased specialization of labor, management, and equipment.

Because minimum efficient scale (MES) is extremely large for commercial aircraft, Boeing only has two production facilities in the United States. By contrast, MES is very small in concrete manufacturing. So there are thousands of concrete makers in the United States.

TERMS AND CONCEPTS

economic cost
explicit costs
implicit costs
accounting profit
normal profit
economic profit
short run
long run
total product (TP)
marginal product (MP)
average product (AP)
law of diminishing returns
fixed costs
variable costs
total cost
average fixed cost (AFC)
average variable cost (AVC)
average total cost (ATC)
marginal cost (MC)
economies of scale
diseconomies of scale
constant returns to scale
minimum efficient scale (MES)
natural monopoly

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Distinguish between explicit and implicit costs, giving examples of each. What are some explicit and implicit costs of attending college? **LO9.1**
2. Distinguish between accounting profit, economic profit, and normal profit. Does accounting profit or economic profit determine how entrepreneurs allocate resources between different business ventures? Explain. **LO9.1**
3. Complete the table below by calculating marginal product and average product.

Inputs of Labor	Total Product	Marginal Product	Average Product
0	0		
		______	
1	15		______
		______	
2	34		______
		______	
3	51		______
		______	
4	65		______
		______	
5	74		______
		______	
6	80		______
		______	
7	83		______
		______	
8	82		______

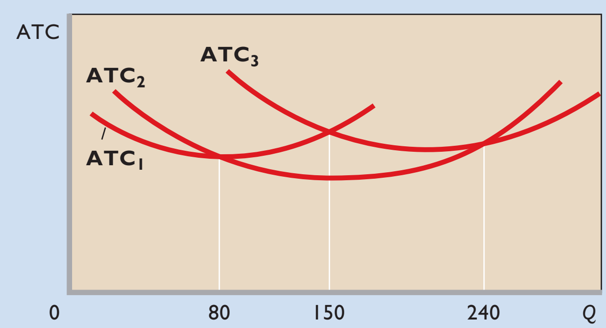
Plot the total, marginal, and average products and explain in detail the relationship between each pair of curves. Explain why marginal product first rises, then declines, and ultimately becomes negative. What bearing does the law of diminishing returns have on short-run costs? Be specific. "When marginal product is rising, marginal cost is falling. And when marginal product is diminishing, marginal cost is rising." Illustrate and explain graphically. **LO9.2**
4. Why can the distinction between fixed costs and variable costs be made in the short run? Classify the following as fixed or variable costs: advertising expenditures, fuel, interest on company-issued bonds, shipping charges, payments for raw materials, real estate taxes, executive salaries, insurance premiums, wage payments, depreciation and obsolescence charges, sales taxes, and rental payments on leased office machinery. "There are no fixed costs in the long run; all costs are variable." Explain. **LO9.3**
5. List several fixed and variable costs associated with owning and operating an automobile. Suppose you are considering whether to drive your car or fly 1,000 miles to Florida for spring break. Which costs—fixed, variable, or both—would you take into account in making your decision? Would any implicit costs be relevant? Explain. **LO9.3**
6. Use the concepts of economies and diseconomies of scale to explain the shape of a firm's long-run ATC curve. What is the concept of minimum efficient scale? What bearing can

the shape of the long-run ATC curve have on the structure of an industry? **LO9.4**

7. **LAST WORD** Does additive manufacturing rely on economies of scale to deliver low costs? What are two ways in which additive manufacturing lowers costs? Besides what's written in the book, might there be another reason to expect 3-D blueprints to be inexpensive? (Hint: Think in terms of supply and demand.)

REVIEW QUESTIONS

1. Linda sells 100 bottles of homemade ketchup for $10 each. The cost of the ingredients, the bottles, and the labels was $700. In addition, it took her 20 hours to make the ketchup and to do so she took time off from a job that paid her $20 per hour. Linda's accounting profit is ____________ while her economic profit is ____________. **LO9.1**
   a. $700; $400
   b. $300; $100
   c. $300; negative $100
   d. $1,000; negative $1,100
2. Which of the following are short-run and which are long-run adjustments? **LO9.1**
   a. Wendy's builds a new restaurant.
   b. Harley-Davidson Corporation hires 200 more production workers.
   c. A farmer increases the amount of fertilizer used on his corn crop.
   d. An Alcoa aluminum plant adds a third shift of workers.
3. A firm has fixed costs of $60 and variable costs as indicated in the table at the bottom of this page. Complete the table and check your calculations by referring to problem 4 at the end of Chapter 10. **LO9.3**
   a. Graph total fixed cost, total variable cost, and total cost. Explain how the law of diminishing returns influences the shapes of the variable-cost and total-cost curves.
   b. Graph AFC, AVC, ATC, and MC. Explain the derivation and shape of each of these four curves and their relationships to one another. Specifically, explain in nontechnical terms why the MC curve intersects both the AVC and the ATC curves at their minimum points.
   c. Explain how the location of each curve graphed in question 3b would be altered if (1) total fixed cost had been $100 rather than $60 and (2) total variable cost had been $10 less at each level of output.
4. Indicate how each of the following would shift the (1) marginal-cost curve, (2) average-variable-cost curve, (3) average-fixed-cost curve, and (4) average-total-cost curve of a manufacturing firm. In each case specify the direction of the shift. **LO9.3**
   a. A reduction in business property taxes.
   b. An increase in the nominal wages of production workers.
   c. A decrease in the price of electricity.
   d. An increase in insurance rates on plant and equipment.
   e. An increase in transportation costs.
5. True or false. The U shape of the long-run ATC curve is the result of diminishing returns. **LO9.4**
6. Suppose a firm has only three possible plant-size options, represented by the ATC curves shown in the accompanying figure. What plant size will the firm choose in producing (*a*) 50, (*b*) 130, (*c*) 160, and (*d*) 250 units of output? Draw the firm's long-run average-cost curve on the diagram and describe this curve. **LO9.4**



Total Product	Total Fixed Cost	Total Variable Cost	Total Cost	Average Fixed Cost	Average Variable Cost	Average Total Cost	Marginal Cost
0	$______	$ 0	$______			$______	
1	______	45	______	$______	$______	______	$______
2	______	85	______	______	______	______	______
3	______	120	______	______	______	______	______
4	______	150	______	______	______	______	______
5	______	185	______	______	______	______	______
6	______	225	______	______	______	______	______
7	______	270	______	______	______	______	______
8	______	325	______	______	______	______	______
9	______	390	______	______	______	______	______
10	______	465	______	______	______	______	______

PROBLEMS

1. Gomez runs a small pottery firm. He hires one helper at \$12,000 per year, pays annual rent of \$5,000 for his shop, and spends \$20,000 per year on materials. He has \$40,000 of his own funds invested in equipment (pottery wheels, kilns, and so forth) that could earn him \$4,000 per year if alternatively invested. He has been offered \$15,000 per year to work as a potter for a competitor. He estimates his entrepreneurial talents are worth \$3,000 per year. Total annual revenue from pottery sales is \$72,000. Calculate the accounting profit and the economic profit for Gomez's pottery firm. **LO9.1**
2. Imagine you have some workers and some handheld computers that you can use to take inventory at a warehouse. There are diminishing returns to taking inventory. If one worker uses one computer, he can inventory 100 items per hour. Two workers sharing a computer can together inventory 150 items per hour. Three workers sharing a computer can together inventory 160 items per hour. And four or more workers sharing a computer can together inventory fewer than 160 items per hour. Computers cost \$100 each and you must pay each worker \$25 per hour. If you assign one worker per computer, what is the cost of inventorying a single item? What if you assign two workers per computer? Three? How many workers per computer should you assign if you wish to minimize the cost of inventorying a single item? **LO9.2**
3. You are a newspaper publisher. You are in the middle of a one-year rental contract for your factory that requires you to pay \$500,000 per month, and you have contractual labor obligations of \$1 million per month that you can't get out of. You also have a marginal printing cost of \$0.25 per paper as well as a marginal delivery cost of \$0.10 per paper. If sales fall by 20 percent from 1 million papers per month to 800,000 papers per month, what happens to the AFC per paper, the MC per paper, and the minimum amount that you must charge to break even on these costs? **LO9.3**
4. There are economies of scale in ranching, especially with regard to fencing land. Suppose that barbed-wire fencing costs \$10,000 per mile to set up. How much would it cost to fence a single property whose area is one square mile if that property also happens to be perfectly square, with sides that are each one mile long? How much would it cost to fence exactly four such properties, which together would contain four square miles of area? Now, consider how much it would cost to fence in four square miles of ranch land if, instead, it comes as a single large square that is two miles long on each side. Which is more costly—fencing in the four, one-square-mile properties or the single four-square-mile property? **LO9.4**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **10**

Pure Competition in the Short Run

Learning Objectives

LO10.1 **Give the names and summarize the main characteristics of the four basic market models.**

LO10.2 **List the conditions required for purely competitive markets.**

LO10.3 **Explain how demand is seen by a purely competitive seller.**

LO10.4 **Convey how purely competitive firms can use the total-revenue–total-cost approach to maximize profits or minimize losses in the short run.**

LO10.5 **Explain how purely competitive firms can use the marginal-revenue–marginal-cost approach to maximize profits or minimize losses in the short run.**

LO10.6 **Explain why a competitive firm's marginal cost curve is the same as its supply curve.**

In Chapter 6 we examined the relationship between product demand and total revenue, and in Chapter 9 we discussed production costs. Now we want to connect revenues and costs to see how a business decides what price to charge and how much output to produce. A firm's decisions concerning price and production depend greatly on the character of the industry in which it is operating. There is no "average" or "typical" industry. At one extreme is an industry in which a single producer dominates the market; at the other extreme are industries in which thousands of firms each produces a tiny fraction of market supply. Between these extremes are many other types of industries.

Since we cannot examine each industry individually, we will focus on four basic *models* of **market structure.** Together, these models will help you understand how price and output are determined in the many product markets in the economy. They also will help you evaluate the efficiency or inefficiency of those markets. Finally, these four models will provide a crucial background for assessing public policies (such as antitrust policy) relating to certain firms and industries.

Four Market Models

LO10.1 Give the names and summarize the main characteristics of the four basic market models.

Economists group industries into four distinct market structures: pure competition, pure monopoly, monopolistic competition, and oligopoly. These four market models differ in several respects: the number of firms in the industry, whether those firms produce a standardized product or try to differentiate their products from those of other firms, and how easy or how difficult it is for firms to enter the industry.

Very briefly the four models are as follows:

- **Pure competition** involves a very large number of firms producing a standardized product (that is, a product like cotton, for which each producer's output is virtually identical to that of every other producer.) New firms can enter or exit the industry very easily.
- **Pure monopoly** is a market structure in which one firm is the sole seller of a product or service (for example, a local electric utility). Since the entry of additional firms is blocked, one firm constitutes the entire industry. The pure monopolist produces a single unique product, so product differentiation is not an issue.
- **Monopolistic competition** is characterized by a relatively large number of sellers producing differentiated products (clothing, furniture, books). Present in this model is widespread *nonprice competition*, a selling strategy in which a firm does not try to distinguish its product on the basis of price but instead on attributes like design and workmanship (an approach called *product differentiation*). Either entry to or exit from monopolistically competitive industries is quite easy.
- **Oligopoly** involves only a few sellers of a standardized or differentiated product, so each firm is affected by the decisions of its rivals and must take those decisions into account in determining its own price and output.

Table 10.1 summarizes the characteristics of the four models for easy comparison and later reference. In discussing these market models, we will occasionally distinguish the

TABLE 10.1 Characteristics of the Four Basic Market Models

	Market Model			
Characteristic	**Pure Competition**	**Monopolistic Competition**	**Oligopoly**	**Pure Monopoly**
Number of firms	A very large number	Many	Few	One
Type of product	Standardized	Differentiated	Standardized or differentiated	Unique; no close substitutes
Control over price	None	Some, but within rather narrow limits	Limited by mutual interdependence; considerable with collusion	Considerable
Conditions of entry	Very easy, no obstacles	Relatively easy	Significant obstacles	Blocked
Nonprice competition	None	Considerable emphasis on advertising, brand names, trademarks	Typically a great deal, particularly with product differentiation	Mostly public relations advertising
Examples	Agriculture	Retail trade, dresses, shoes	Steel, automobiles, farm implements, many household appliances	Local utilities

characteristics of *pure competition* from those of the three other basic market structures, which together we will designate as **imperfect competition.**

Pure Competition: Characteristics and Occurrence

LO10.2 List the conditions required for purely competitive markets.

Although pure competition is relatively rare in the real world, this market model is highly relevant to several industries. In particular, we can learn much about markets for agricultural goods, fish products, foreign exchange, basic metals, and stock shares by studying the pure-competition model. Also, pure competition is a meaningful starting point for any discussion of price and output determination. Moreover, the operation of a purely competitive economy provides a standard, or norm, for evaluating the efficiency of the real-world economy.

Let's take a fuller look at pure competition, the focus of the remainder of this chapter:

- ***Very large numbers*** A basic feature of a purely competitive market is the presence of a large number of independently acting sellers, often offering their products in large national or international markets. Examples: markets for farm commodities, the stock market, and the foreign exchange market.
- ***Standardized product*** Purely competitive firms produce a standardized (identical or homogeneous) product. As long as the price is the same, consumers will be indifferent about which seller to buy the product from. Buyers view the products of firms B, C, D, and E as perfect substitutes for the product of firm A. Because purely competitive firms sell standardized products, they make no attempt to differentiate their products and do not engage in other forms of nonprice competition.
- ***"Price takers"*** In a purely competitive market, individual firms do not exert control over product price. Each firm produces such a small fraction of total output that increasing or decreasing its output will not perceptibly influence total supply or, therefore, product price. In short, the competitive firm is a **price taker:** It cannot change market price; it can only adjust to it. That means that the individual competitive producer is at the mercy of the market. Asking a price higher than the market price would be futile. Consumers will not buy from firm A at $2.05 when its 9,999 competitors are selling an identical product, and therefore a perfect substitute, at $2 per unit. Conversely, because firm A can sell as much as it chooses at $2 per unit, it has no reason to charge a lower price, say, $1.95. Doing that would shrink its profit.
- ***Free entry and exit*** New firms can freely enter and existing firms can freely leave purely competitive industries. No significant legal, technological, financial, or other obstacles prohibit new firms from selling their output in any competitive market.

Demand as Seen by a Purely Competitive Seller

LO10.3 Explain how demand is seen by a purely competitive seller.

We begin by examining demand from a purely competitive seller's viewpoint to see how it affects revenue. This seller might be a wheat farmer, a strawberry grower, a sheep rancher, a foreign-currency broker, or some other pure competitor. Because each purely competitive firm offers only a negligible fraction of total market supply, it must accept the price determined by the market; it is a price taker, not a price maker.

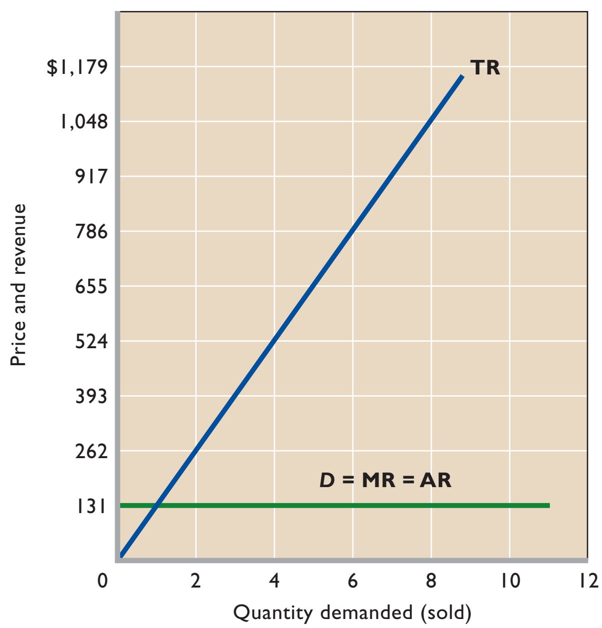
Perfectly Elastic Demand

The demand schedule faced by the *individual firm* in a purely competitive industry is perfectly elastic at the market price, as demonstrated in Figure 10.1. As shown in column 1 of the table in Figure 10.1, the market price is $131. The firm represented cannot obtain a higher price by restricting its output, nor does it need to lower its price to increase its sales volume. Columns 1 and 2 show that the firm can produce and sell as many or as few units as it likes at the market price of $131.

We are *not* saying that *market* demand is perfectly elastic in a competitive market. Rather, market demand graphs as a downsloping curve. An entire industry (all firms producing a particular product) can affect price by changing industry output. For example, all firms, acting independently but simultaneously, can increase price by reducing output. But the individual competitive firm cannot do that because its output represents such a small fraction of its industry's total output. For the individual competitive firm, the market price is therefore a fixed value at which it can sell as many or as few units as it cares to. Graphically, this implies that the individual competitive firm's demand curve will plot as a straight, horizontal line such as *D* in Figure 10.1.

FIGURE 10.1 A purely competitive firm's demand and revenue curves. The demand curve (*D*) of a purely competitive firm is a horizontal line (perfectly elastic) because the firm can sell as much output as it wants at the market price (here, $131). Because each additional unit sold increases total revenue by the amount of the price, the firm's total-revenue (TR) curve is a straight upsloping line and its marginal-revenue (MR) curve coincides with the firm's demand curve. The average-revenue (AR) curve also coincides with the demand curve.

Firm's Demand Schedule		Firm's Revenue Data	
(1) Product Price (*P*) (Average Revenue)	(2) Quantity Demanded (Q)	(3) Total Revenue (TR), (1) × (2)	(4) Marginal Revenue (MR)
$131	0	$ 0	
			$131
131	1	131	
			131
131	2	262	
			131
131	3	393	
			131
131	4	524	
			131
131	5	655	
			131
131	6	786	
			131
131	7	917	
			131
131	8	1048	
			131
131	9	1179	
			131
131	10	1310	



Average, Total, and Marginal Revenue

The firm's demand schedule is also its average-revenue schedule. Price per unit to the purchaser is also revenue per unit, or average revenue, to the seller. To say that all buyers must pay $131 per unit is to say that the revenue per unit, or **average revenue** received by the seller, is $131. Price and average revenue are the same thing.

The **total revenue** for each sales level is found by multiplying price by the corresponding quantity the firm can sell. (Column 1 multiplied by column 2 in the table in Figure 10.1 yields column 3.) In this case, total revenue increases by a constant amount, $131, for each additional unit of sales. Each unit sold adds exactly its constant price—no more or no less—to total revenue.

When a firm is pondering a change in its output, it will consider how its total revenue will change as a result. **Marginal revenue** is the change in total revenue (or the extra revenue) that results from selling one more unit of output. In column 3 of the table in Figure 10.1, total revenue is zero when zero units are sold. The first unit of output sold increases total revenue from zero to $131, so marginal revenue for that unit is $131. The second unit sold increases total revenue from $131 to $262, and marginal revenue is again $131. Note in column 4 that marginal revenue is a constant $131, as is price. *In pure competition, marginal revenue and price are equal.*

Figure 10.1 shows the purely competitive firm's total-revenue, demand, marginal-revenue, and average-revenue curves. Total revenue (TR) is a straight line that slopes upward to the right. Its slope is constant because each extra unit of sales increases TR by $131. The demand curve (*D*) is horizontal, indicating perfect price elasticity. The marginal-revenue (MR) curve coincides with the demand curve because the product price (and hence MR) is constant. The average revenue (AR) curve equals price and therefore also coincides with the demand curve.

QUICK REVIEW 10.1

- In a purely competitive industry a large number of firms produce a standardized product and there are no significant barriers to entry.
- The demand seen by a purely competitive firm is perfectly elastic—horizontal on a graph—at the market price.
- Marginal revenue and average revenue for a purely competitive firm coincide with the firm's demand curve; total revenue rises by the product price for each additional unit sold.

Profit Maximization in the Short Run: Total-Revenue–Total-Cost Approach

LO10.4 Convey how purely competitive firms can use the total-revenue–total-cost approach to maximize profits or minimize losses in the short run.

Because the purely competitive firm is a price taker, it cannot attempt to maximize its profit by raising or lowering the price it charges. With its price set by supply and demand in the overall market, the only variable that the firm can control is its output. As a result, the purely competitive firm attempts to maximize its economic profit (or minimize its economic loss) by adjusting its *output*. And, in the short run, the firm has a fixed plant. Thus it can adjust its output only through changes in the amount of variable resources (materials, labor) it uses. It adjusts its variable resources to achieve the output level that maximizes its profit or minimizes its loss.

There are two ways to determine the level of output at which a competitive firm will realize maximum profit or minimum loss. One method is to compare total revenue and total cost; the other is to compare marginal revenue and marginal cost. Both approaches apply to all firms, whether they are pure competitors, pure monopolists, monopolistic competitors, or oligopolists.[1]

We begin by examining profit maximization using the total-revenue–total-cost approach. Confronted with the market price of its product, the competitive producer will ask three questions: (1) Should we produce this product? (2) If so, in what amount? (3) What economic profit (or loss) will we realize?

Let's demonstrate how a pure competitor answers these questions, given a particular set of cost data and a specific market price. Our cost data are already familiar because they are the fixed-cost, variable-cost, and total-cost data in Table 9.2, repeated in columns 1 to 4 of the table in Figure 10.2. (Recall that these data reflect explicit and implicit costs, including a normal profit.) Assuming that the market price is $131, the total revenue for each output level is found by multiplying output (total product) by price. Total-revenue data are in column 5. Then in column 6 we find the profit or loss at each output level by subtracting total cost, TC (column 4), from total revenue, TR (column 5).

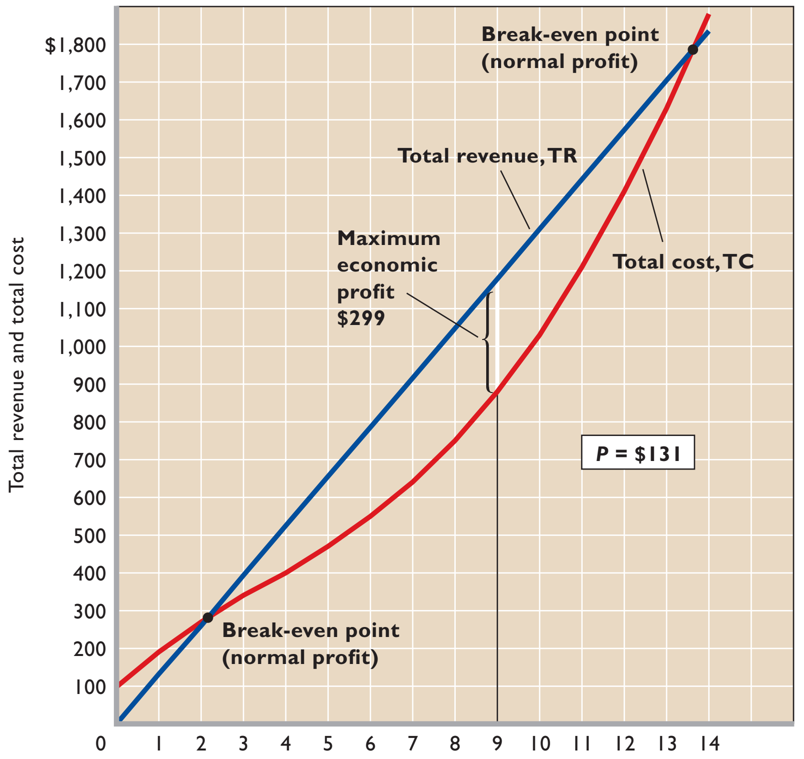
WORKED PROBLEMS

W10.1
Profit maximization: TR−TC approach

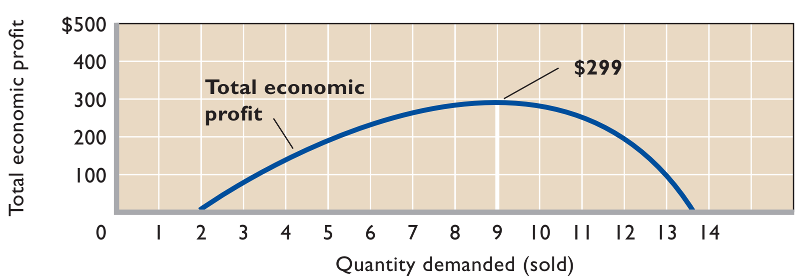
Should the firm produce? Definitely. It can obtain a profit by doing so. How much should it produce? Nine units. Column 6 tells us that this is the output at which total economic profit is at a maximum. What economic profit (or loss) will it realize? A $299 economic profit—the difference between total revenue ($1,179) and total cost ($880).

Figure 10.2a compares total revenue and total cost graphically for this profit-maximizing case. Observe again that the total-revenue curve for a purely competitive firm is a straight line (Figure 10.1). Total cost increases with

[1]To make sure you understand these two approaches, we will apply both of them to output determination under pure competition. But since we want to emphasize the marginal approach, we will limit our graphical application of the total-revenue approach to a situation where the firm maximizes profits. We will then use the marginal approach to examine three cases: profit maximization, loss minimization, and shutdown.



(a)
Profit-maximizing case



(b)
Total economic profit

FIGURE 10.2 Total-revenue–total-cost approach to profit maximization for a purely competitive firm. (a) The firm's profit is maximized at that output (9 units) where total revenue, TR, exceeds total cost, TC, by the maximum amount. (b) The vertical distance between TR and TC in (a) is plotted as a total-economic-profit curve. Maximum economic profit is $299 at 9 units of output.

				PRICE: $131	
(1) Total Product (Output) (*Q*)	(2) Total Fixed Cost (TFC)	(3) Total Variable Cost (TVC)	(4) Total Cost (TC)	(5) Total Revenue (TR)	(6) Profit (+) or Loss (−)
0	$100	$ 0	$ 100	$ 0	$−100
1	100	90	190	131	−59
2	100	170	270	262	−8
3	100	240	340	393	+53
4	100	300	400	524	+124
5	100	370	470	655	+185
6	100	450	550	786	+236
7	100	540	640	917	+277
8	100	650	750	1,048	+298
9	100	780	**880**	**1,179**	**+299**
10	100	930	1,030	1,310	+280

output because more production requires more resources. But the *rate* of increase in total cost varies with the efficiency of the firm, which in turn varies with the amount of variable inputs that are being combined with the firm's current amount of capital (which is fixed in the short run). Stated slightly differently, the cost data reflect Chapter 9's law of diminishing returns. From zero to four units of output, total cost increases at a decreasing rate as the firm temporarily experiences increasing returns. At higher levels of output, however, efficiency falls as crowding causes diminishing returns to set in. Once that happens, the firm's total cost increases at an increasing rate because each additional unit of input yields less output than the previous unit.

Total revenue and total cost are equal where the two curves in Figure 10.2a intersect (at roughly 2 units of output). Total revenue covers all costs (including a normal profit, which is included in the cost curve), but there is no economic profit. For this reason economists call this output a **break-even point:** an output at which a firm makes a *normal profit* but not an economic profit. If we extended the data beyond 10 units of output, another break-even point would occur where total cost catches up with total revenue, somewhere between 13 and 14 units of output in Figure 10.2a. Any output within the two break-even points identified in the figure will yield an economic profit. The firm achieves maximum profit, however, where the vertical distance between the total-revenue and total-cost curves is greatest. For our particular data, this is at 9 units of output, where maximum profit is $299.

The profit-maximizing output is easier to see in Figure 10.2b, where total profit is graphed for each level of output. Where the total-revenue and total-cost curves intersect in Figure 10.2a, economic profit is zero, as shown by the total-profit line in Figure 10.2b. Where the vertical distance between TR and TC is greatest in the upper graph, economic profit is at its peak ($299), as shown in the lower graph. This firm will choose to produce 9 units since that output maximizes its profit.

Profit Maximization in the Short Run: Marginal-Revenue–Marginal-Cost Approach

LO10.5 Explain how purely competitive firms can use the marginal-revenue–marginal-cost approach to maximize profits or minimize losses in the short run.

In the second approach, the firm compares the amounts that each *additional* unit of output would add to total revenue and to total cost. In other words, the firm compares the *marginal revenue* (MR) and the *marginal cost* (MC) of each successive unit of output. Assuming that producing is preferable to shutting down, the firm should produce any unit of output whose marginal revenue exceeds its marginal cost because the firm would gain more in revenue from selling that unit than it would add to its costs by producing it. Conversely, if the marginal cost of a unit of output exceeds its marginal revenue, the firm should not produce that unit. Producing it would add more to costs than to revenue, and profit would decline or loss would increase.

In the initial stages of production, where output is relatively low, marginal revenue will usually (but not always) exceed marginal cost. So it is profitable to produce through this range of output. But at later stages of production, where output is relatively high, rising marginal costs will exceed marginal revenue. Obviously, a profit-maximizing firm will want to avoid output levels in that range. Separating these two production ranges is a unique point at which marginal revenue equals marginal cost. This point is the key to the output-determining rule: *In the short run, the firm will maximize profit or minimize loss by producing the output at which marginal revenue equals marginal cost (as long as producing is preferable to shutting down).* This profit-maximizing guide is known as the **MR = MC rule.**

Keep in mind these features of the MR = MC rule:

- For most sets of MR and MC data, MR and MC will be precisely equal at a fractional level of output. In such instances the firm should produce the last complete unit of output for which MR exceeds MC.
- As noted, the rule applies only if producing is preferable to shutting down. We will show shortly that if marginal revenue does not equal or exceed average variable cost, the firm will shut down rather than produce the amount of output at which MR = MC.
- The rule is an accurate guide to profit maximization for all firms whether they are purely competitive, monopolistic, monopolistically competitive, or oligopolistic.
- The rule can be restated as P = MC when applied to a purely competitive firm. Because the demand schedule faced by a competitive seller is perfectly elastic at the going market price, product price and marginal revenue are equal. So under pure competition (and only under pure competition) we may substitute P for MR in the rule: When producing is preferable to shutting down, the competitive firm that wants to maximize its profit or minimize its loss should produce at that point where price equals marginal cost ($P = MC$).

Now let's apply the MR = MC rule or, because we are considering pure competition, the *P* = MC rule, first using the same price as used in our total-revenue–total-cost approach to profit maximization. Then, by considering other prices, we will demonstrate two additional cases: loss minimization and shutdown. It is crucial that you understand the MR = MC analysis that follows since it reappears in Chapters 12, 13, and 14.

Profit-Maximizing Case

The first five columns of the table in **Figure 10.3 (Key Graph)** reproduce the AFC, AVC, ATC, and MC data derived for our product in Table 9.2. It is the marginal-cost data of column 5 that we will compare with price (equals marginal revenue) for each unit of output. Suppose first that the market price, and therefore marginal revenue, is \$131, as shown in column 6.

What is the profit-maximizing output? Every unit of output up to and including the ninth unit represents greater marginal revenue than marginal cost of output. Each of the first 9 units therefore adds to the firm's profit and should be produced. The tenth unit, however, should not be produced. It would add more to cost (\$150) than to revenue (\$131). So 9 units is the profit-maximizing output.

The economic profit realized by producing 9 units can be calculated by subtracting total cost from total revenue. Multiplying price (\$131) by output (9), we find that total revenue is \$1,179. From the average-total-cost data in column 4, we see that ATC is \$97.78 at 9 units of output. Multiplying \$97.78 by 9 gives us total cost of \$880.[2] The difference of \$299 (= \$1,179 − \$880) is the economic profit. Clearly, this firm will prefer to operate rather than shut down.

Perhaps an easier way to calculate the economic profit is to use this simple equation, in which *A* is average total cost:

$$\text{Profit} = (P - A) \times Q$$

So by subtracting the average total cost (\$97.78) from the product price (\$131), we obtain a per-unit profit of \$33.22. Multiplying that amount by 9 units of output, we determine that the profit is \$299. Take some time now to verify the numbers in column 7. You will find that any output other than that which adheres to the MR = MC rule will mean either profits below \$299 or losses.

WORKED PROBLEMS

W10.2
Profit maximization: MR = MC approach

The graph in Figure 10.3 shows price (= MR) and marginal cost graphically. Price equals marginal cost at the profit-maximizing output of 9 units. There the per-unit economic profit is $P - A$, where P is the market price and A is the average total cost for an output of 9 units. The total economic profit is $9 \times (P - A)$, shown by the green rectangular area.

Note that the firm wants to maximize its total profit, not its per-unit profit. Per-unit profit is greatest at 7 units of output, where price exceeds average total cost by \$39.57 (= \$131 − \$91.43). But by producing only 7 units, the firm would be forgoing the production of 2 additional units of output that would clearly contribute to total profit. The firm is happy to accept lower per-unit profits for additional units of output because they nonetheless add to total profit.

Loss-Minimizing Case

Now let's assume that the market price is \$81 rather than \$131. Should the firm still produce? If so, how much? And what will be the resulting profit or loss? The answers, respectively, are "Yes," "Six units," and "A loss of \$64."

The first five columns of the table in Figure 10.4 are the same as the first five columns of the table in Figure 10.3. But column 6 of the table in Figure 10.4 shows the new price (equal to MR), \$81. Comparing columns 5 and 6, we find that the first unit of output adds \$90 to total cost but only \$81 to total revenue. One might conclude: "Don't produce—close down!" But that would be hasty. Remember that in the very early stages of production, marginal product is low, making marginal cost unusually high. The price–marginal cost relationship improves with increased production. For units 2 through 6, price exceeds marginal cost. Each of these 5 units adds more to revenue than to cost, and as shown in column 7, they decrease the total loss. Together they more than compensate for the "loss" taken on the first unit. Beyond 6 units, however, MC exceeds MR (= *P*). The firm should therefore produce 6 units. In general, the profit-seeking producer should always compare marginal revenue (or price under pure competition) with the rising portion of the marginal-cost schedule or curve.

Will production be profitable? No, because at 6 units of output the average total cost of \$91.67 exceeds the price of \$81 by \$10.67 per unit. If we multiply that by the 6 units of output, we find the firm's total loss is \$64. Alternatively, comparing the total revenue of \$486 (= 6 × \$81) with the total cost of \$550 (= 6 × \$91.67), we see again that the firm's loss is \$64.

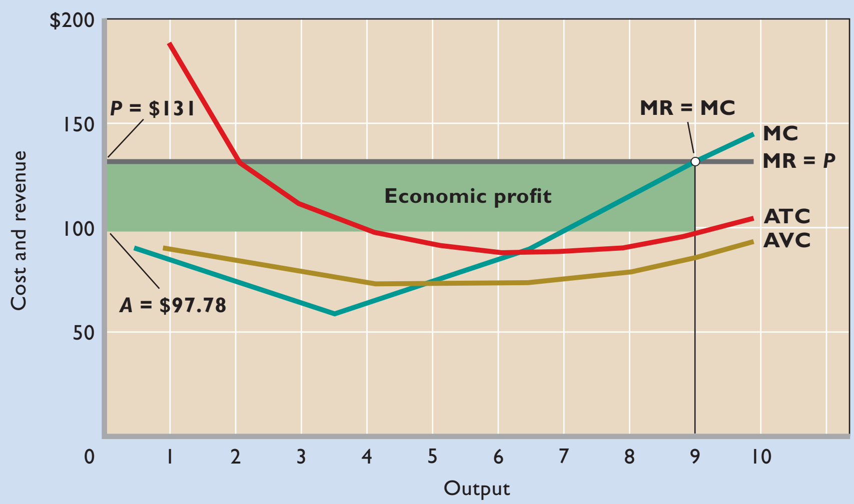
---

[2]Most of the unit-cost data are rounded figures. Therefore, economic profits calculated from them will typically vary by a few cents from the profits determined in the total-revenue–total-cost approach. Here we simply ignore the few-cents differentials to make our answers consistent with the results of the total-revenue–total-cost approach.

KEY GRAPH

FIGURE 10.3 Short-run profit maximization for a purely competitive firm. The MR = MC output enables the purely competitive firm to maximize profits or to minimize losses. In this case MR (= *P* in pure competition) and MC are equal at an output *Q* of 9 units. There, *P* exceeds the average total cost *A* = \$97.78, so the firm realizes an economic profit of *P* − *A* per unit. The total economic profit is represented by the green rectangle and is 9 × (*P* − *A*).

(1) Total Product (Output)	(2) Average Fixed Cost (AFC)	(3) Average Variable Cost (AVC)	(4) Average Total Cost (ATC)	(5) Marginal Cost (MC)	(6) Price = Marginal Revenue (MR)	(7) Total Economic Profit (+) or Loss (−)
0						\$−100
1	\$100.00	\$90.00	\$190.00	\$ 90	\$131	−59
2	50.00	85.00	135.00	80	131	−8
3	33.33	80.00	113.33	70	131	+53
4	25.00	75.00	100.00	60	131	+124
5	20.00	74.00	94.00	70	131	+185
6	16.67	75.00	91.67	80	131	+236
7	14.29	77.14	91.43	90	131	+277
8	12.50	81.25	93.75	110	131	+298
9	**11.11**	**86.67**	**97.78**	**130**	**131**	**+299**
10	10.00	93.00	103.00	150	131	+280

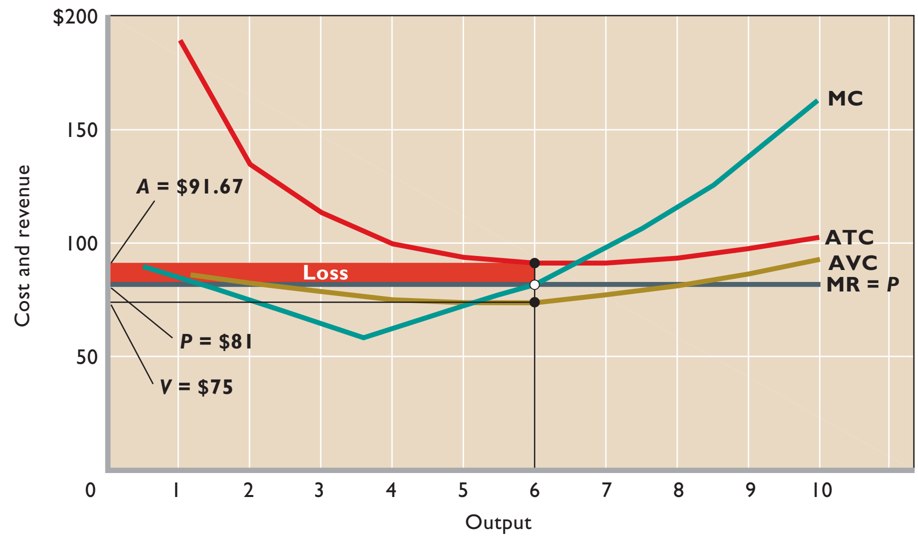


QUICK QUIZ FOR FIGURE 10.3

1. Curve MR is horizontal because:
   a. product price falls as output increases.
   b. the law of diminishing marginal utility is at work.
   c. the market demand for this product is perfectly elastic.
   d. the firm is a price taker.
2. At a price of \$131 and 7 units of output:
   a. MR exceeds MC, and the firm should expand its output.
   b. total revenue is less than total cost.
   c. AVC exceeds ATC.
   d. the firm would earn only a normal profit.
3. In maximizing profits at 9 units of output, this firm is adhering to which of the following decision rules?
   a. Produce where MR exceeds MC by the greatest amount.
   b. Produce where *P* exceeds ATC by the greatest amount.
   c. Produce where total revenue exceeds total cost by the greatest amount.
   d. Produce where average fixed costs are zero.
4. Suppose price declined from \$131 to \$100. This firm's:
   a. marginal-cost curve would shift downward.
   b. economic profit would fall to zero.
   c. profit-maximizing output would decline.
   d. total cost would fall by more than its total revenue.

Answers: 1. d; 2. a; 3. c; 4. c

FIGURE 10.4 Short-run loss minimization for a purely competitive firm. If price *P* exceeds the minimum AVC (here, $74 at *Q* = 5) but is less than ATC, the MR = MC output (here, 6 units) will permit the firm to minimize its losses. In this instance the loss is *A* − *P* per unit, where *A* is the average total cost at 6 units of output. The total loss is shown by the red area and is equal to 6 × (*A* − *P*).



					Loss-Minimizing Case		Shutdown Case	
(1) Total Product (Output)	(2) Average Fixed Cost (AFC)	(3) Average Variable Cost (AVC)	(4) Average Total Cost (ATC)	(5) Marginal Cost (MC)	(6) $81 Price = Marginal Revenue (MR)	(7) Profit (+) or Loss (−), $81 Price	(8) $71 Price = Marginal Revenue (MR)	(9) Profit (+) or Loss (−), $71 Price
0						$−100		**$−100**
1	$100.00	$90.00	$190.00	$ 90	$81	−109	$71	−119
2	50.00	85.00	135.00	80	81	−108	71	−128
3	33.33	80.00	113.33	70	81	−97	71	−127
4	25.00	75.00	100.00	60	81	−76	71	−116
5	20.00	74.00	94.00	70	81	−65	71	−115
6	**16.67**	**75.00**	**91.67**	**80**	**81**	**−64**	71	−124
7	14.29	77.14	91.43	90	81	−73	71	−143
8	12.50	81.25	93.75	110	81	−102	71	−182
9	11.11	86.67	97.78	130	81	−151	71	−241
10	10.00	93.00	103.00	150	81	−220	71	−320

Then why produce? Because this loss is less than the firm's $100 of fixed costs, which is the $100 loss the firm would incur in the short run by closing down. The firm receives enough revenue per unit ($81) to cover its average variable costs of $75 and also provide $6 per unit, or a total of $36, to apply against fixed costs. Therefore, the firm's loss is only $64 (= $100 − $36), not $100.

This loss-minimizing case is illustrated in the graph in Figure 10.4. Wherever price *P* exceeds average variable cost AVC but is less than ATC, the firm can pay part, but not all, of its fixed costs by producing. The loss is minimized by producing the output at which MC = MR (here, 6 units). At that output, each unit contributes $P - V$ to covering fixed cost, where V is the AVC at 6 units of output. The per-unit loss is $A - P = \$10.67$, and the total loss is $6 \times (A - P)$, or $64, as shown by the red area.

Shutdown Case

Suppose now that the market yields a price of only $71. Should the firm produce? No, because at every output level the firm's average variable cost is greater than the price (compare columns 3 and 8 of the table in Figure 10.4). The smallest loss it can incur by producing is greater

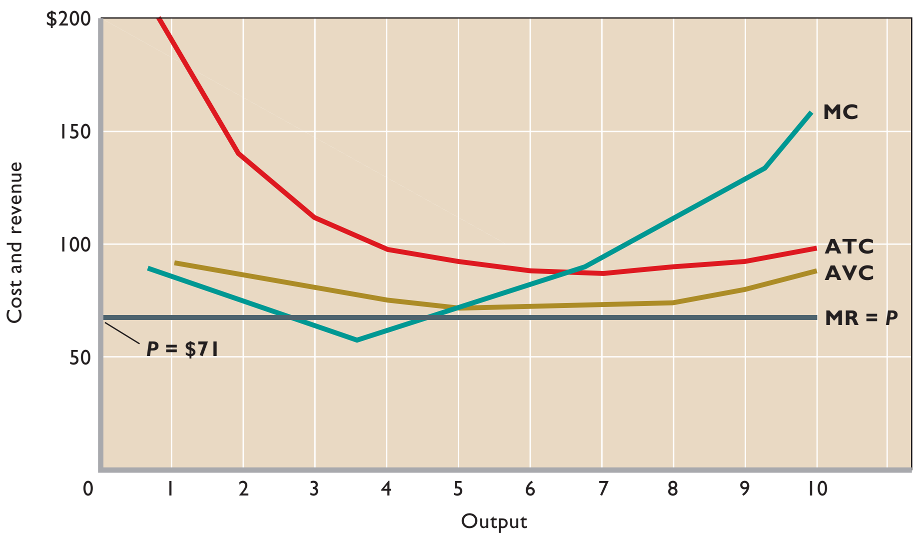


FIGURE 10.5 The short-run shutdown case for a purely competitive firm. If price P falls below the minimum AVC (here, $74 at $Q = 5$), the competitive firm will minimize its losses in the short run by shutting down. There is no level of output at which the firm can produce and incur a loss smaller than its total fixed cost.

than the $100 fixed cost it will lose by shutting down (as shown by column 9). The best action is to shut down.

You can see this shutdown situation in Figure 10.5. Price comes closest to covering average variable costs at the MR (= P) = MC output of 5 units. But even here, price or revenue per unit would fall short of average variable cost by $3 (= $74 − $71). By producing at the MR (= P) = MC output, the firm would lose its $100 worth of fixed cost plus $15 ($3 of variable cost on each of the 5 units), for a total loss of $115. This compares unfavorably with the $100 fixed-cost loss the firm would incur by shutting down and producing no output. So it will make sense for the firm to shut down rather than produce at a $71 price—or at any price less than the minimum average variable cost of $74.

The shutdown case reminds us of the qualifier to our MR (= P) = MC rule. A competitive firm will maximize profit or minimize loss in the short run by producing that output at which MR (= P) = MC, *provided that market price exceeds minimum average variable cost.*

QUICK REVIEW 10.2

- A firm will choose to produce if it can at least break even and generate a normal profit.
- Profit is maximized, or loss minimized, at the output at which marginal revenue (or price in pure competition) equals marginal cost, provided that price exceeds variable cost.
- If the market price is below the minimum average variable cost, the firm will minimize its losses by shutting down.

Marginal Cost and Short-Run Supply

LO10.6 Explain why a competitive firm's marginal cost curve is the same as its supply curve.

In the preceding section we simply selected three different prices and asked what quantity the profit-seeking competitive firm, faced with certain costs, would choose to offer in the market at each price. This set of product prices and corresponding quantities supplied constitutes part of the supply schedule for the competitive firm.

Table 10.2 summarizes the supply schedule data for those three prices ($131, $81, and $71) and four others. This table confirms the direct relationship between product price and quantity supplied that we identified in Chapter 3. Note first that the firm will not produce at price $61 or $71 because both are less than the $74 minimum AVC. Then note that quantity supplied increases as price increases. Observe finally that economic profit is higher at higher prices.

TABLE 10.2 The Supply Schedule of a Competitive Firm Confronted with the Cost Data in the table in Figure 10.3

Price	Quantity Supplied	Maximum Profit (+) or Minimum Loss (−)
$151	10	$+480
131	9	+299
111	8	+138
91	7	−3
81	6	−64
71	0	−100
61	0	−100

KEY GRAPH

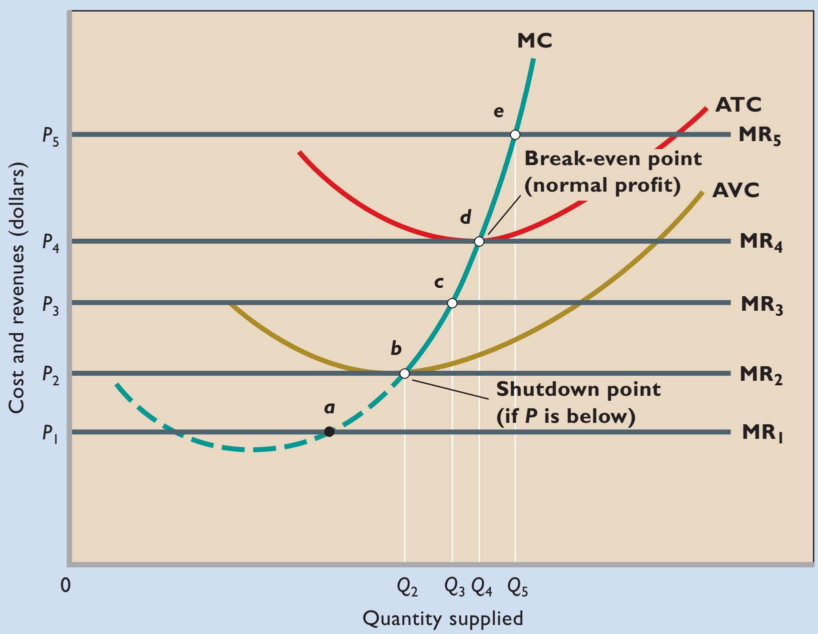


FIGURE 10.6 The $P = MC$ rule and the competitive firm's short-run supply curve. Application of the $P = MC$ rule, as modified by the shutdown case, reveals that the (solid) segment of the firm's MC curve that lies above AVC is the firm's short-run supply curve. More specifically, at price P_1, $P = MC$ at point *a*, but the firm will produce no output because P_1 is less than minimum AVC. At price P_2 the firm will operate at point *b*, where it produces Q_2 units and incurs a loss equal to its total fixed cost. At P_3 it operates at point *c*, where output is Q_3 and the loss is less than total fixed cost. With the price of P_4, the firm operates at point *d*; in this case the firm earns a normal profit because at output Q_4 price equals ATC. At price P_5 the firm operates at point *e* and maximizes its economic profit by producing Q_5 units.

QUICK QUIZ FOR FIGURE 10.6

1. Which of the following might increase product price from P_3 to P_5?
   - **a.** An improvement in production technology.
   - **b.** A decline in the price of a substitute good.
   - **c.** An increase in the price of a complementary good.
   - **d.** Rising incomes if the product is a normal good.
2. An increase in price from P_3 to P_5 would:
   - **a.** shift this firm's MC curve to the right.
   - **b.** mean that MR_5 exceeds MC at Q_3 units, inducing the firm to expand output to Q_5.
   - **c.** decrease this firm's average variable costs.
   - **d.** enable this firm to obtain a normal, but not an economic, profit.
3. At P_4:
   - **a.** this firm has no economic profit.
   - **b.** this firm will earn only a normal profit and thus will shut down.
   - **c.** MR_4 will be less than MC at the profit-maximizing output.
   - **d.** the profit-maximizing output will be Q_5.
4. Suppose P_4 is \$10, P_5 is \$15, Q_4 is 8 units, and Q_5 is 10 units. This firm's:
   - **a.** supply curve is elastic over the Q_4–Q_5 range of output.
   - **b.** supply curve is inelastic over the Q_4–Q_5 range of output.
   - **c.** total revenue will decline if price rises from P_4 to P_5.
   - **d.** marginal-cost curve will shift downward if price falls from P_5 to P_4.

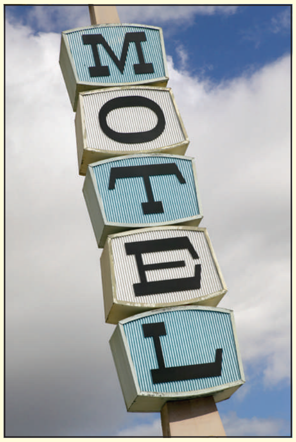
Answers: 1. d; 2. b; 3. a; 4. b

Generalized Depiction

Figure 10.6 (Key Graph) generalizes the MR = MC rule and the relationship between short-run production costs and the firm's supply behavior. The ATC, AVC, and MC curves are shown, along with several marginal-revenue lines drawn at possible market prices. Let's observe quantity supplied at each of these prices:

- Price P_1 is below the firm's minimum average variable cost, so at this price the firm won't operate at all. Quantity supplied will be zero, as it will be at all other prices below P_2.
- Price P_2 is just equal to the minimum average variable cost. The firm will supply Q_2 units of output (where MR_2 = MC) and just cover its total variable cost. Its loss will equal its total fixed cost. (Actually, the firm would be indifferent as to shutting down or supplying Q_2 units of output, but we assume it produces.)

CONSIDER THIS . . .



The "Still There" Motel

Have you ever driven by a poorly maintained business facility and wondered why the owner does not either fix up the property or go out of business? The somewhat surprising reason is that it may be unprofitable to improve the facility yet profitable to continue to operate the business as it deteriorates. Seeing why will aid your understanding of the "stay open or shut down" decision facing firms experiencing declining demand.

Consider the story of the Still There Motel on Old Highway North, Anytown, USA. The owner built the motel on the basis of traffic patterns and competition existing several decades ago. But as interstate highways were built, the motel found itself located on a relatively untraveled stretch of road. Also, it faced severe competition from "chain" motels located much closer to the interstate highway.

As demand and revenue fell, Still There moved from profitability to loss ($P <$ ATC). But at first its room rates and annual revenue were sufficient to cover its total variable costs and contribute some to the payment of fixed costs such as insurance and property taxes ($P >$ AVC). By staying open, Still There lost less than it would have if it shut down. But since its total revenue did not cover its total costs (or $P <$ ATC), the owner realized that something must be done in the long run. The owner decided to lower total costs by reducing annual maintenance. In effect, the owner opted to allow the motel to deteriorate as a way of temporarily regaining profitability.

This renewed profitability of Still There cannot last because in time no further reduction of maintenance costs will be possible. The deterioration of the motel structure will produce even lower room rates, and therefore even less total revenue. The owner of Still There knows that sooner or later total revenue will again fall below total cost (or P will again fall below ATC), even with an annual maintenance expense of zero. When that occurs, the owner will close down the business, tear down the structure, and sell the vacant property. But, in the meantime, the motel is still there—open, deteriorating, and profitable.

- At price P_3 the firm will supply Q_3 units of output to minimize its short-run losses. At any of the other prices between P_2 and P_4 the firm will also minimize its losses by producing and supplying the quantity at which MR (= P) = MC.
- The firm will just break even at price P_4. There it will supply Q_4 units of output (where MR_4 = MC), earning a normal profit but not an economic profit. Total revenue will just cover total cost, including a normal profit, because the revenue per unit ($MR_4 = P_4$) and the total cost per unit (ATC) are the same.
- At price P_5 the firm will realize an economic profit by producing and supplying Q_5 units of output. In fact, at any price above P_4 the firm will obtain economic profit by producing to the point where MR (= P) = MC.

Note that each of the MR (= P) = MC intersection points labeled *b*, *c*, *d*, and *e* in Figure 10.6 indicates a possible product price (on the vertical axis) and the corresponding quantity that the firm would supply at that price (on the horizontal axis). Thus, points such as these are on the upsloping supply curve of the competitive firm. Note, too, that quantity supplied would be zero at any price below the minimum average variable cost (AVC). *We can conclude that the portion of the firm's marginal-cost curve lying above its average-variable-cost curve is its short-run supply curve.* In Figure 10.6, the solid segment of the marginal-cost curve MC is this firm's **short-run supply curve.** It tells us the amount of output the firm will supply at each price in a series of prices.

Table 10.3 summarizes the MR = MC approach to determining the competitive firm's profit-maximizing output level. It also shows the equivalent analysis in terms of total revenue and total cost.

Diminishing Returns, Production Costs, and Product Supply

We have now identified the links between the law of diminishing returns (Chapter 9), production costs, and product supply in the short run. Because of the law of

TABLE 10.3 Output Determination in Pure Competition in the Short Run

Question	Answer
Should this firm produce?	Yes, if price is equal to, or greater than, minimum average variable cost. This means that the firm is profitable or that its losses are less than its fixed cost.
What quantity should this firm produce?	Produce where MR (= P) = MC; there, profit is maximized (TR exceeds TC by a maximum amount) or loss is minimized.
Will production result in economic profit?	Yes, if price exceeds average total cost (so that TR exceeds TC). No, if average total cost exceeds price (so that TC exceeds TR).

diminishing returns, marginal costs eventually rise as more units of output are produced. And because marginal costs rise with output, a purely competitive firm must get successively higher prices to motivate it to produce additional units of output.

Viewed alternatively, higher product prices and marginal revenue encourage a purely competitive firm to expand output. As its output increases, the firm's marginal costs rise as a result of the law of diminishing returns. At some now greater output, the higher MC equals the new product price and MR. Profit once again is maximized, but at a greater total amount. Quantity supplied has increased in direct response to an increase in product price and the desire to maximize profit.

TABLE 10.4 **Firm and Market Supply and Market Demand**

(1) Quantity Supplied, Single Firm	(2) Total Quantity Supplied, 1,000 Firms	(3) Product Price	(4) Total Quantity Demanded
10	10,000	$151	4,000
9	9,000	131	6,000
8	8,000	111	8,000
7	7,000	91	9,000
6	6,000	81	11,000
0	0	71	13,000
0	0	61	16,000

Changes in Supply

In Chapter 9 we saw that changes in such factors as the prices of variable inputs or in technology will alter costs and shift the marginal-cost or short-run supply curve to a new location. All else equal, for example, a wage increase would increase marginal cost and shift the supply curve in Figure 10.6 upward as viewed from the horizontal axis (leftward as viewed from the vertical axis). That is, supply would decrease. Similarly, technological progress that increases the productivity of labor would reduce marginal cost and shift the marginal-cost or supply curve downward as viewed from the horizontal axis (rightward as viewed from the vertical axis). This represents an increase in supply.

Firm and Industry: Equilibrium Price

In the preceding section we established the competitive firm's short-run supply curve by applying the MR (= *P*) = MC rule. But which of the various possible prices will actually be the market equilibrium price?

From Chapter 3 we know that the market equilibrium price will be the price at which the total quantity supplied of the product equals the total quantity demanded. So to determine the equilibrium price, we first need to obtain a total supply schedule and a total demand schedule. We find the total supply schedule by assuming a particular number of firms in the industry and supposing that each firm has the same individual supply schedule as the firm represented in Figure 10.6. Then we sum the quantities supplied at each price level to obtain the total (or market) supply schedule. Columns 1 and 3 in Table 10.4 repeat the supply schedule for the individual competitive firm, as derived in Table 10.2. Suppose 1,000 firms compete in this industry, all having the same total and unit costs as the single firm we discussed. This lets us calculate the market supply schedule (columns 2 and 3) by multiplying the quantity-supplied figures of the single firm (column 1) by 1,000.

Market Price and Profits To determine the equilibrium price and output, these total-supply data must be compared with total-demand data. Let's assume that total demand is as shown in columns 3 and 4 in Table 10.4. By comparing the total quantity supplied and the total quantity demanded at the seven possible prices, we determine that the equilibrium price is $111 and the equilibrium quantity is 8,000 units for the industry—8 units for each of the 1,000 identical firms.

Will these conditions of market supply and demand make this a profitable or unprofitable industry? Multiplying product price ($111) by output (8 units), we find that the total revenue of each firm is $888. The total cost is $750, found by looking at column 4 of the table in Figure 10.2. The $138 difference is the economic profit of each firm. For the industry, total economic profit is $138,000. This, then, is a profitable industry.

Another way of calculating economic profit is to determine per-unit profit by subtracting average total cost ($93.75) from product price ($111) and multiplying the difference (per-unit profit of $17.25) by the firm's equilibrium level of output (8). Again we obtain an economic profit of $138 per firm and $138,000 for the industry.

Figure 10.7 shows this analysis graphically. The individual supply curves of each of the 1,000 identical firms—one of which is shown as s = MC in Figure 10.7a—are summed horizontally to get the total-supply curve $S = \Sigma$MC of Figure 10.7b. With total-demand curve *D*, it yields the equilibrium price $111 and equilibrium quantity (for the industry) 8,000 units. This equilibrium price is given and unalterable to the individual firm; that is, each firm's demand curve is perfectly elastic at the equilibrium price, as indicated

LAST WORD

Fixed Costs: Digging Yourself Out of a Hole

For Firms Facing Losses Due to Fixed Costs, Shutting Down in the Short Run Does Not Mean Shutting Down Forever.

A firm with fixed costs starts each month standing at the bottom of a deep financial hole. The depth of that "money pit" is equal to the dollar value of all the payments that the firm is legally obligated to make even if it is producing nothing. These fixed costs include contractually guaranteed salaries, interest payments on loans, and equipment rental fees that are locked-in by long-term contracts. As the firm stands at the bottom of this fixed-cost financial hole and stares upward looking for a way out, it has to ask itself the following question: Will producing output make the hole even deeper?

Naturally, the firm hopes that producing output will generate positive cash flows that will offset its fixed costs and start filling in the hole. If those positive flows are large enough, they may completely offset the firm's fixed costs and fill up the hole, thereby allowing the firm to break even. And if they are just a bit larger, they will not only fill up the hole but also accumulate a nice little pile of profits above ground.

But those are just the firm's hopes. The firm's reality may be quite unpleasant. In particular, the firm may be facing a situation in which producing output would make its financial situation worse rather than better. As explained in this chapter, if the price of the firm's output falls too low, then producing output will yield cash flows that are negative rather than positive because revenues will be less than variable costs. If that happens, producing output will lose money for the firm so that the firm would be

by *d* in Figure 10.7a. Because the individual firm is a price taker, the marginal-revenue curve coincides with the firm's demand curve *d*. This $111 price exceeds the average total cost at the firm's equilibrium MR = MC output of 8 units, so the firm earns an economic profit represented by the green area in Figure 10.7a.

Assuming no changes in costs or market demand, these diagrams reveal a genuine equilibrium in the short run.

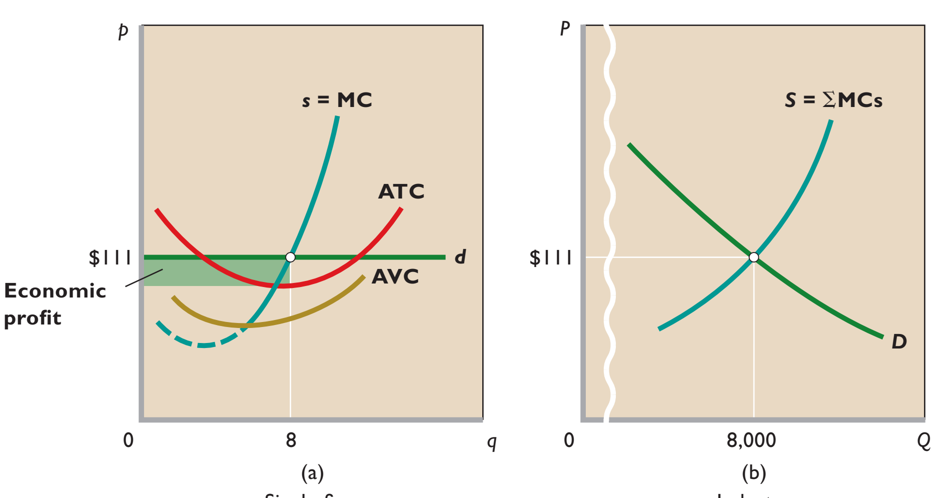


FIGURE 10.7 Short-run competitive equilibrium for (a) a firm and (b) the industry. The horizontal sum of the 1,000 firms' individual supply curves (*s*) determines the industry supply curve (*S*). Given industry demand (*D*), the short-run equilibrium price and output for the industry are $111 and 8,000 units. Taking the equilibrium price as given, the individual firm establishes its profit-maximizing output at 8 units and, in this case, realizes the economic profit represented by the green area.

better off shutting down production rather than producing output. By shutting down, it will lose only its fixed costs. By shutting down, its financial hole won't get even deeper.

A crucial thing to understand, however, is that the low prices that cause firms to shut down production are often temporary—so that shutdowns are also often temporary. Just because a firm shuts down at a given moment to prevent its financial hole from getting any deeper does not mean that the firm will go out of business forever. To the contrary, many industries are characterized by firms that regularly switch production on and off depending upon the market price they can get for their output and, consequently, whether producing output will generate positive or negative cash flows.

Oil production is a good example. Different wells have different variable production costs. If the price of oil drops below a given well's variable costs, then it would be better to halt production on that well and just lose the value of its fixed costs rather than pumping oil whose variable cost exceeds the revenue that it generates when sold.

Seasonal resorts are another good example of turning production on and off depending on the price. The demand for hotel rooms near ski resorts in New Hampshire, for instance, is much higher during the winter ski season than it is during the summer. As a result, the market price of hotel rooms falls so low during the summer that many inns and resorts close during the warmer months. They have all sorts of fixed costs, but it makes more sense for them to shut down rather than remain open because operating in the summer would cost more in variable costs than it would generate in revenues. Better to lose only their fixed costs.

Numerous other examples of temporary shutdowns occur during recessions, the occasional economy-wide economic slowdowns during which demand declines for nearly all goods and services. The 2007–2009 recession in the United States, for instance, saw many manufacturing companies temporarily shut down and mothball their production facilities. The recession witnessed the mothballing of electric generating plants, factories that make fiber optic cable, automobile factories, chemical plants, textile mills, and even the plant in McIntosh, Alabama, that makes the artificial sweetener Splenda. Many other firms also shut down production to wait out the recession—so many, in fact, that there was a mini-boom for consulting firms that specialized in helping firms mothball their factories (the main problem being how to properly store idle machinery so that it will work again when it is eventually brought back into service).

Firms that mothball factories or equipment during a recession do so expecting to eventually turn them back on. But the lengths of recessions vary, as do the specific circumstances of individual firms. So while many firms shut down in the short run with the expectation of reopening as soon as their particular business conditions improve, sometimes their business conditions do not improve. Sometimes the only way to terminate fixed costs is to terminate the firm.

No shortages or surpluses occur in the market to cause price or total quantity to change. Nor can any firm in the industry increase its profit by altering its output. Note, too, that higher unit and marginal costs, on the one hand, or weaker market demand, on the other, could change the situation so that Figure 10.7a resembles Figure 10.4 or Figure 10.5.

Firm versus Industry Figure 10.7 underscores a point made earlier: Product price is a given fact to the *individual* competitive firm, but the supply plans of all competitive producers *as a group* are a basic determinant of product price. If we recall the fallacy of composition (Last Word, Chapter 1), we find there is no inconsistency here. Although one firm, supplying a negligible fraction of total supply, cannot affect price, the sum of the supply curves of all the firms in the industry constitutes the industry supply curve, and that curve does have an important bearing on price.

WORKED PROBLEMS

W10.3
Short-run competitive equilibrium

QUICK REVIEW 10.3

- A competitive firm's short-run supply curve is the portion of its marginal cost (MC) curve that lies above its average variable cost (AVC) curve.
- If price P is greater than minimum average variable cost, the firm will produce the amount of output where MR ($= P$) $=$ MC in order to either maximize its profit (if price exceeds minimum ATC) or minimize its loss (if price lies between minimum AVC and minimum ATC).
- Market supply in a competitive industry is the horizontal sum of the individual supply curves of all of the firms in the industry. The market equilibrium price is determined by where the industry's market supply curve intersects the industry's market demand curve.

SUMMARY

LO10.1 Give the names and summarize the main characteristics of the four basic market models.

Economists group industries into four models based on their market structures: (*a*) pure competition, (*b*) pure monopoly, (*c*) monopolistic competition, and (*d*) oligopoly.

LO10.2 List the conditions required for purely competitive markets.

A purely competitive industry consists of a large number of independent firms producing a standardized product. Pure competition assumes that firms and resources are mobile among different industries.

LO10.3 Explain how demand is seen by a purely competitive seller.

In a competitive industry, no single firm can influence market price. This means that the firm's demand curve is perfectly elastic and price equals marginal revenue.

LO10.4 Convey how purely competitive firms can use the total-revenue–total-cost approach to maximize profits or minimize losses in the short run.

We can analyze short-run profit maximization by a competitive firm by comparing total revenue and total cost or by applying marginal analysis. A firm maximizes its short-run profit by producing the output at which total revenue exceeds total cost by the greatest amount.

LO10.5 Explain how purely competitive firms can use the marginal-revenue–marginal-cost approach to maximize profits or minimize losses in the short run.

Provided price exceeds minimum average variable cost, a competitive firm maximizes profit or minimizes loss in the short run by producing the output at which price or marginal revenue equals marginal cost.

If price is less than minimum average variable cost, a competitive firm minimizes its loss by shutting down. If price is greater than average variable cost but is less than average total cost, a competitive firm minimizes its loss by producing the P = MC amount of output. If price also exceeds average total cost, the firm maximizes its economic profit at the P = MC amount of output.

LO10.6 Explain why a competitive firm's marginal cost curve is the same as its supply curve.

Applying the MR (= P) = MC rule at various possible market prices leads to the conclusion that the segment of the firm's short-run marginal-cost curve that lies above the firm's average-variable-cost curve is its short-run supply curve.

A competitive firm shuts down production at least temporarily if price is less than minimum average variable cost because, in those situations, producing any amount of output will always result in variable costs exceeding revenues. Shutting down therefore results in a smaller loss because the firm will lose only its fixed cost, whereas, if it operated, it would lose its fixed cost plus whatever money is lost due to variable costs exceeding revenues.

Competitive firms choose to operate rather than shut down whenever price is greater than average variable cost but less than average total cost because, in those situations, revenues will always exceed variable costs. The amount by which revenues exceed variable costs can be used to help pay down some of the firm's fixed costs. Thus, the firm loses less money by operating (and paying down some of its fixed costs) than it would if it shut down (in which case it would suffer a loss equal to the full amount of its fixed costs).

TERMS AND CONCEPTS

market structure
pure competition
pure monopoly
monopolistic competition
oligopoly
imperfect competition
price taker
average revenue
total revenue
marginal revenue
break-even point
MR = MC rule
short-run supply curve

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Briefly state the basic characteristics of pure competition, pure monopoly, monopolistic competition, and oligopoly. Under which of these market classifications does each of the following most accurately fit? (*a*) a supermarket in your hometown; (*b*) the steel industry; (*c*) a Kansas wheat farm; (*d*) the commercial bank in which you or your family has an account; (*e*) the automobile industry. In each case, justify your classification. **LO10.1**
2. Strictly speaking, pure competition is relatively rare. Then why study it? **LO10.2**

3. "Even if a firm is losing money, it may be better to stay in business in the short run." Is this statement ever true? Under what condition(s)? **LO10.5**
4. Consider a firm that has no fixed costs and that is currently losing money. Are there any situations in which it would want to stay open for business in the short run? If a firm has no fixed costs, is it sensible to speak of the firm distinguishing between the short run and the long run? **LO10.5**
5. Why is the equality of marginal revenue and marginal cost essential for profit maximization in all market structures? Explain why price can be substituted for marginal revenue in the MR = MC rule when an industry is purely competitive. **LO10.5**
6. "That segment of a competitive firm's marginal-cost curve that lies above its average-variable-cost curve constitutes the short-run supply curve for the firm." Explain using a graph and words. **LO10.6**
7. **LAST WORD** If a firm's current revenues are less than its current variable costs, when should it shut down? If the firm decides to shut down, should we expect that decision to be final? Explain using an example that is not in the book.

REVIEW QUESTIONS

1. Suppose that the paper clip industry is perfectly competitive. Also assume that the market price for paper clips is 2 cents per paper clip. The demand curve faced by each firm in the industry is: **LO10.3**
   a. A horizontal line at 2 cents per paper clip.
   b. A vertical line at 2 cents per paper clip.
   c. The same as the market demand curve for paper clips.
   d. Always higher than the firm's MC curve.
2. Use the demand schedule below to determine total revenue and marginal revenue for each possible level of sales: **LO10.3**
   a. What can you conclude about the structure of the industry in which this firm is operating? Explain.

Product Price	Quantity Demanded	Total Revenue	Marginal Revenue
$2	0	$______	
			$______
2	1	______	
			______
2	2	______	
			______
2	3	______	
			______
2	4	______	
			______
2	5	______	

   b. Graph the demand, total-revenue, and marginal-revenue curves for this firm.
   c. Why do the demand and marginal-revenue curves coincide?
   d. "Marginal revenue is the change in total revenue associated with additional units of output." Explain verbally and graphically, using the data in the table.
3. A purely competitive firm whose goal is to maximize profit will choose to produce the amount of output at which: **LO10.4**
   a. TR and TC are equal.
   b. TR exceeds TC by as much as possible.
   c. TC exceeds TR by as much as possible.
   d. none of the above.
4. If it is possible for a perfectly competitive firm to do better financially by producing rather than shutting down, then it should produce the amount of output at which: **LO10.5**
   a. MR < MC.
   b. MR = MC.
   c. MR > MC.
   d. none of the above.
5. A perfectly competitive firm that makes car batteries has a fixed cost of $10,000 per month. The market price at which it can sell its output is $100 per battery. The firm's minimum AVC is $105 per battery. The firm is currently producing 500 batteries a month (the output level at which MR = MC). This firm is making a ____________ and should ____________ production. **LO10.5**
   a. profit; increase
   b. profit; shut down
   c. loss; increase
   d. loss; shut down
6. Consider a profit-maximizing firm in a competitive industry. For each of the following situations, indicate whether the firm should shut down production or produce where MR = MC. **LO10.5**
   a. P < minimum AVC.
   b. P > minimum ATC.
   c. Minimum AVC < P < minimum ATC.

PROBLEMS

1. A purely competitive firm finds that the market price for its product is $20. It has a fixed cost of $100 and a variable cost of $10 per unit for the first 50 units and then $25 per unit for all successive units. Does price exceed average variable cost for the first 50 units? What about for the first 100 units? What is the marginal cost per unit for the first 50 units? What about for units 51 and higher? For each of the first 50 units, does MR exceed MC? What about for units 51 and higher? What output level will yield the largest possible profit for this purely competitive firm? (Hint: Draw a graph similar to Figure 10.2 using data for this firm.) **LO10.5**

2. A purely competitive wheat farmer can sell any wheat he grows for $10 per bushel. His five acres of land show diminishing returns because some are better suited for wheat production than others. The first acre can produce 1,000 bushels of wheat, the second acre 900, the third 800, and so on. Draw a table with multiple columns to help you answer the following questions. How many bushels will each of the farmer's five acres produce? How much revenue will each acre generate? What are the TR and MR for each acre? If the marginal cost of planting and harvesting an acre is $7,000 per acre for each of the five acres, how many acres should the farmer plant and harvest? **LO10.5**
3. Karen runs a print shop that makes posters for large companies. It is a very competitive business. The market price is currently $1 per poster. She has fixed costs of $250. Her variable costs are $1,000 for the first thousand posters, $800 for the second thousand, and then $750 for each additional thousand posters. What is her AFC per poster (not per thousand!) if she prints 1,000 posters? 2,000? 10,000? What is her ATC per poster if she prints 1,000? 2,000? 10,000? If the market price fell to 70 cents per poster, would there be *any* output level at which Karen would *not* shut down production immediately? **LO10.5**
4. Assume that the cost data in the following table are for a purely competitive producer: **LO10.5**

Total Product	Average Fixed Cost	Average Variable Cost	Average Total Cost	Marginal Cost
0				
1	$60.00	$45.00	$105.00	$45
2	30.00	42.50	72.50	40
3	20.00	40.00	60.00	35
4	15.00	37.50	52.50	30
5	12.00	37.00	49.00	35
6	10.00	37.50	47.50	40
7	8.57	38.57	47.14	45
8	7.50	40.63	48.13	55
9	6.67	43.33	50.00	65
10	6.00	46.50	52.50	75

a. At a product price of $56, will this firm produce in the short run? If it is preferable to produce, what will be the profit-maximizing or loss-minimizing output? What economic profit or loss will the firm realize per unit of output?

b. Answer the questions of 4a assuming product price is $41.

c. Answer the questions of 4a assuming product price is $32.

d. In the table below, complete the short-run supply schedule for the firm (columns 1 and 2) and indicate the profit or loss incurred at each output (column 3).

(1) Price	(2) Quantity Supplied, Single Firm	(3) Profit (+) or Loss (−)	(4) Quantity Supplied 1,500 Firms
$26	______	$______	______
32	______	______	______
38	______	______	______
41	______	______	______
46	______	______	______
56	______	______	______
66	______	______	______

e. Now assume that there are 1,500 identical firms in this competitive industry; that is, there are 1,500 firms, each of which has the cost data shown in the table. Complete the industry supply schedule (column 4).

f. Suppose the market demand data for the product are as follows:

Price	Total Quantity Demanded
$26	17,000
32	15,000
38	13,500
41	12,000
46	10,500
56	9,500
66	8,000

What will be the equilibrium price? What will be the equilibrium output for the industry? For each firm? What will profit or loss be per unit? Per firm? Will this industry expand or contract in the long run?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **11**

Pure Competition in the Long Run

Learning Objectives

LO11.1 **Explain how the long run differs from the short run in pure competition.**

LO11.2 **Describe how profits and losses drive the long-run adjustment process of pure competition.**

LO11.3 **Explain the differences between constant-cost, increasing-cost, and decreasing-cost industries.**

LO11.4 **Show how long-run equilibrium in pure competition produces an efficient allocation of resources.**

LO11.5 **Discuss creative destruction and the profit incentives for innovation.**

The previous chapter discussed how pure competition operates in the short run, the time period during which the individual firms in an industry are stuck with their current plant sizes and fixed-cost commitments. As you know, pure competitors shut down production if prices are too low or, if prices are high enough, produce where MR = MC to minimize their losses or maximize their profits. Whether they make a profit or a loss depends on how high the market price is relative to their costs.

That being said, profits and losses cannot be the end of the pure competition story because one of the key characteristics of pure competition is the freedom of firms to enter or exit the industry. We know from Chapter 2 that profits attract entry and losses prompt exit.

In this chapter, we are keenly interested in how entry and exit relate to allocative and productive efficiency. We are also interested in how continuing competition leads to new products and new business methods replacing older products and older business methods through a process aptly referred to as *creative destruction.*

The Long Run in Pure Competition

LO11.1 Explain how the long run differs from the short run in pure competition.

The entry and exit of firms in our market models can only take place in the long run. In the short run, the industry is composed of a specific number of firms, each with a plant size that is fixed and unalterable in the short run. Firms may shut down in the sense that they can produce zero units of output in the short run, but they do not have sufficient time to liquidate their assets and go out of business.

In the long run, by contrast, the firms already in an industry have sufficient time to either expand or contract their capacities. More important, the number of firms in the industry may either increase or decrease as new firms enter or existing firms leave.

The length of time constituting the long run varies substantially by industry, however, so that you should not fix in your mind any specific number of years, months, or days. Instead, focus your attention on the incentives provided by profits and losses for the entry and exit of firms into any purely competitive industry and, later in the chapter, on how those incentives lead to productive and allocative efficiency. The time horizons are far less important than the process by which profits and losses guide business managers toward the efficient use of society's resources.

Profit Maximization in the Long Run

The first part of the pure competition story (Chapter 10) was about profit, loss, and shutdown in the short run. The rest of the story (this chapter) is about entry and exit and their effects on industry size and allocative and productive efficiency in the long run.

To tell the rest of story well, we need to return to our graphical analysis and examine profit maximization by pure competitors in the long run. Several assumptions, none of which affect our conclusions, will keep things simple:

- ***Entry and exit only*** The only long-run adjustment in our graphical analysis is caused by the entry or exit of firms. Moreover, we ignore all short-run adjustments in order to concentrate on the effects of the long-run adjustments.
- ***Identical costs*** All firms in the industry have identical cost curves. This assumption lets us discuss an "average," or "representative," firm, knowing that all other firms in the industry are similarly affected by any long-run adjustments that occur.
- ***Constant-cost industry*** The industry is a constant-cost industry. This means that the entry and exit of firms does not affect resource prices or, consequently, the locations of the average-total-cost curves of individual firms.

The Long-Run Adjustment Process in Pure Competition

LO11.2 Describe how profits and losses drive the long-run adjustment process of pure competition.

The basic conclusion we seek to explain is this: After all long-run adjustments are completed in a purely competitive industry, product price will be exactly equal to, and production will occur at, each firm's minimum average total cost.

This conclusion follows from two basic facts: (1) Firms seek profits and shun losses and (2) under pure competition, firms are free to enter and leave an industry. If market price initially exceeds minimum average total cost, the resulting economic profit will attract new firms to the industry. But this industry expansion will increase supply until price is brought back down to equality with minimum average total cost. Conversely, if price is initially less than minimum average total cost, the resulting loss will cause firms to leave the industry. As they leave, total supply will decline, bringing the price back up to equality with minimum average total cost.

Long-Run Equilibrium

Consider the average firm in a purely competitive industry that is initially in long-run equilibrium. This firm is represented in Figure 11.1a, where MR = MC and price and minimum average total cost are equal at \$50. Economic profit here is zero; the industry is in equilibrium or "at rest" because there is no tendency for firms to enter or to leave. The existing firms are earning normal profits, which means that their accounting profits are equal to those that the owners of these firms could expect

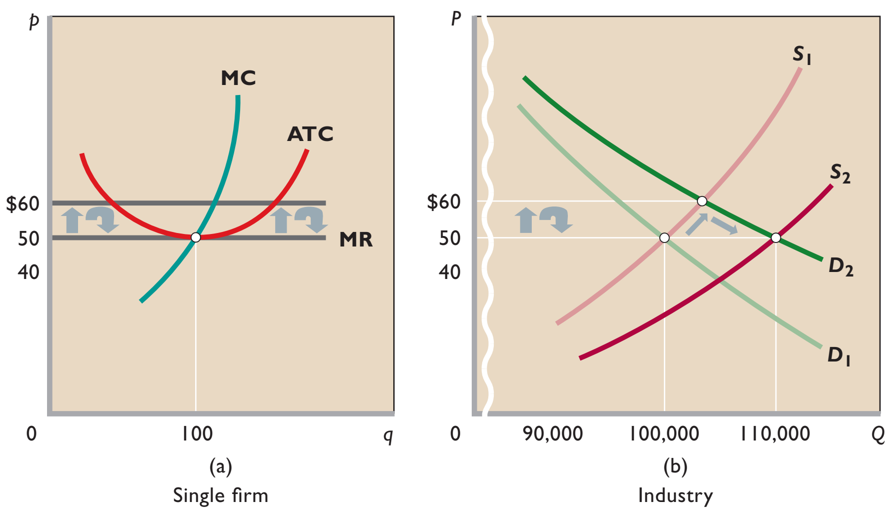


FIGURE 11.1 Temporary profits and the reestablishment of long-run equilibrium in (a) a representative firm and (b) the industry. A favorable shift in demand (D_1 to D_2) will upset the original industry equilibrium and produce economic profits. But those profits will entice new firms to enter the industry, increasing supply (S_1 to S_2) and lowering product price until economic profits are once again zero.

to receive on average in other industries. It is because their current profits are the same as they could expect to earn elsewhere that there is no tendency for firms to enter or leave the industry. The $50 market price is determined in Figure 11.1b by market or industry demand D_1 and supply S_1. (S_1 is a short-run supply curve; we will develop the long-run industry supply curve in our discussion.) And remember that normal profits earned by these firms are considered an opportunity cost and, therefore, are included in the firms' cost curves.

As shown on the quantity axes of the two graphs, equilibrium output in the industry is 100,000 while equilibrium output for the single firm is 100. If all firms in the industry are identical, there must be 1,000 firms (= 100,000/100).

Entry Eliminates Economic Profits Let's upset the long-run equilibrium in Figure 11.1 and see what happens. Suppose a change in consumer tastes increases product demand from D_1 to D_2. Price will rise to $60, as determined at the intersection of D_2 and S_1, and the firm's marginal-revenue curve will shift upward to $60. This $60 price exceeds the firm's average total cost of $50 at output 100, creating an economic profit of $10 per unit. This economic profit will lure new firms into the industry. Some entrants will be newly created firms; others will shift from less prosperous industries.

As firms enter, the market supply of the product increases, pushing the product price below $60. Economic profits persist, and entry continues until short-run supply increases to S_2. Market price falls to $50, as does marginal revenue for the firm. Price and minimum average total cost are again equal at $50. The economic profits caused by the boost in demand have been eliminated, and, as a result, the previous incentive for more firms to enter the industry has disappeared because the firms that remain are earning only a normal profit (zero economic profit). Entry ceases and a new long-run equilibrium is reached.

Observe in Figure 11.1a and 11.1b that total quantity supplied is now 110,000 units and each firm is producing 100 units. Now 1,100 firms rather than the original 1,000 populate the industry. Economic profits have attracted 100 more firms.

Exit Eliminates Losses Now let's consider a shift in the opposite direction. We begin in Figure 11.2b with curves S_1 and D_1 setting the same initial long-run equilibrium situation as in our previous analysis, including the $50 price.

Suppose consumer demand declines from D_1 to D_3. This forces the market price and marginal revenue down to $40, making production unprofitable at the minimum ATC of $50. In time the resulting economic losses will induce firms to leave the industry. Their owners will seek a normal profit elsewhere rather than accept the below-normal profits (losses) now confronting them. As this exodus of firms proceeds, however, industry supply decreases, pushing the price up from $40 toward $50. Losses continue and more firms leave the industry until the supply curve shifts to S_3. Once this happens, price is again $50, just equal to the minimum average total cost. Losses have been eliminated so that the firms that remain are earning only a normal profit (zero economic profit). Since this is no better or worse than entrepreneurs could expect to earn in other business ventures, there is no longer any incentive to exit the industry. Long-run equilibrium is restored.

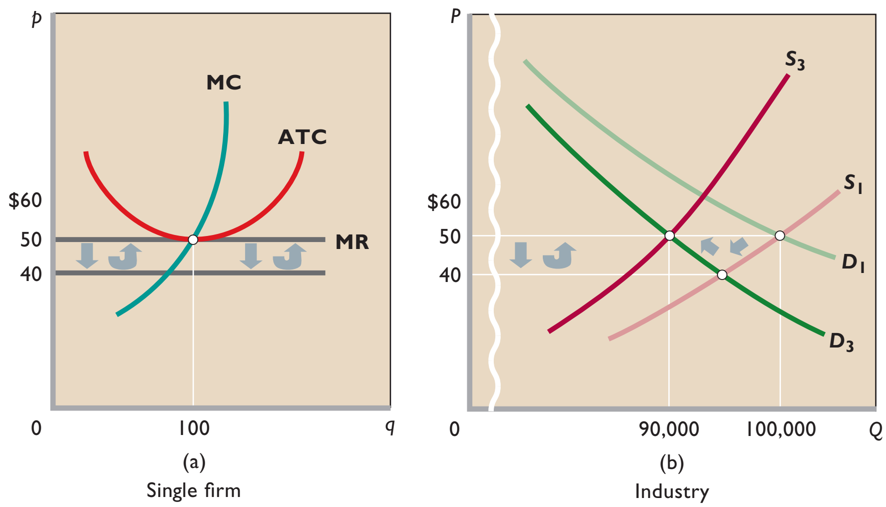


FIGURE 11.2 Temporary losses and the reestablishment of long-run equilibrium in (a) a representative firm and (b) the industry. An unfavorable shift in demand (D_1 to D_3) will upset the original industry equilibrium and produce losses. But those losses will cause firms to leave the industry, decreasing supply (S_1 to S_3) and increasing product price until all losses have disappeared.

In Figure 11.2a and 11.2b, total quantity supplied is now 90,000 units and each firm is producing 100 units. Only 900 firms, not the original 1,000, populate the industry. Losses have forced 100 firms out.

You may have noted that we have sidestepped the question of which firms will leave the industry when losses occur by assuming that all firms have identical cost curves. In the real world, of course, managerial talents differ. Even if resource prices and technology are the same for all firms, less skillfully managed firms tend to incur higher costs and therefore are the first to leave an industry when demand declines. Similarly, firms with less productive labor forces or higher transportation costs will be higher-cost producers and likely candidates to quit an industry when demand decreases.

We have now reached an intermediate goal: Our analysis verifies that competition, reflected in the entry and exit of firms, eliminates economic profits or losses by adjusting price to equal minimum long-run average total cost. In addition, this competition forces firms to select output levels at which average total cost is minimized.

Long-Run Supply Curves

LO11.3 Explain the differences between constant-cost, increasing-cost, and decreasing-cost industries.

Although our analysis has dealt with the long run, we have noted that the market supply curves in Figures 11.1b and 11.2b are short-run curves. What then is the character of the **long-run supply curve** of a competitive industry? Our analysis points us toward an answer. The crucial factor here is the effect, if any, that changes in the number of firms in the industry will have on costs of the individual firms in the industry.

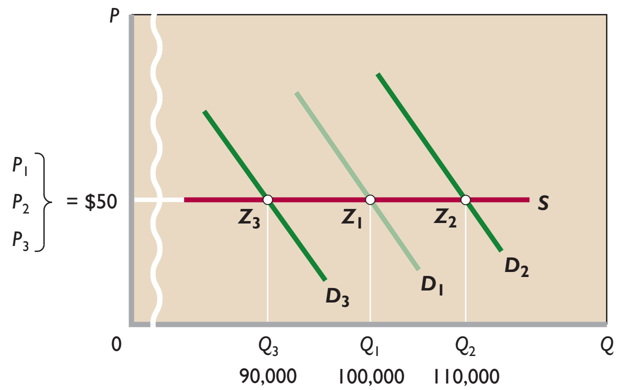
Long-Run Supply for a Constant-Cost Industry

In our analysis of long-run competitive equilibrium we assumed that the industry under discussion was a **constant-cost industry.** This means that industry expansion or contraction will not affect resource prices and therefore production costs. Graphically, it means that the entry or exit of firms does not shift the long-run ATC curves of individual firms. This is the case when the industry's demand for resources is small in relation to the total demand for those resources. Then the industry can expand or contract without significantly affecting resource prices and costs.

What does the long-run supply curve of a constant-cost industry look like? The answer is contained in our previous analysis. There we saw that the entry and exit of firms changes industry output but always brings the product price back to its original level, where it is just equal to the constant minimum ATC. Specifically, we discovered that the industry would supply 90,000, 100,000, or 110,000 units of output, all at a price of \$50 per unit. In other words, the long-run supply curve of a constant-cost industry is perfectly elastic.

This is demonstrated graphically in Figure 11.3, which uses data from Figures 11.1 and 11.2. Suppose industry demand is originally D_1, industry output is Q_1 (100,000 units), and product price is P_1 (\$50). This situation, from Figure 11.1, is one of long-run equilibrium. We saw that when demand increases to D_2, upsetting this equilibrium, the resulting economic profits attract new firms. Because this is a constant-cost industry, entry continues and industry output expands until the price is driven back down to the level of the unchanged minimum ATC. This is at price P_2 (\$50) and output Q_2 (110,000).

FIGURE 11.3 The long-run supply curve for a constant-cost industry. In a constant-cost industry, the entry and exit of firms do not affect resource prices, or, therefore, unit costs. So an increase in demand (D_1 to D_2) raises industry output (Q_1 to Q_2) but not price ($50). Similarly, a decrease in demand (D_1 to D_3) reduces output (Q_1 to Q_3) but not price. Thus, the long-run industry supply curve (S) is horizontal through points Z_1, Z_2, and Z_3.



From Figure 11.2, we saw that a decline in market demand from D_1 to D_3 causes an exit of firms and ultimately restores equilibrium at price P_3 ($50) and output Q_3 (90,000 units). The points Z_1, Z_2, and Z_3 in Figure 11.3 represent these three price-quantity combinations. A line or curve connecting all such points shows the various price-quantity combinations that firms would produce if they had enough time to make all desired adjustments to changes in demand. This line or curve is the industry's long-run supply curve. In a constant-cost industry this curve (straight line) is horizontal, as in Figure 11.3, thus representing perfectly elastic supply.

Long-Run Supply for an Increasing-Cost Industry

Constant-cost industries are a special case. Most industries are **increasing-cost industries,** in which firms' ATC curves shift upward as the industry expands and downward as the industry contracts. Usually, the entry of new firms will increase resource prices, particularly in industries using specialized resources whose long-run supplies do not readily increase in response to increases in resource demand. Higher resource prices result in higher long-run average total costs for all firms in the industry. These higher costs cause upward shifts in each firm's long-run ATC curve.

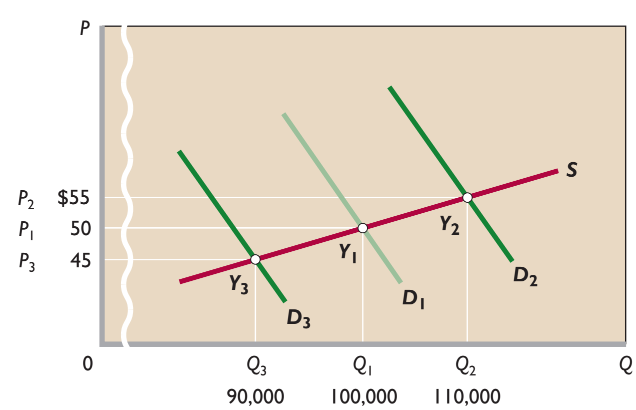
Thus, when an increase in product demand results in economic profits and attracts new firms to an increasing-cost industry, a two-way squeeze works to eliminate those profits. As before, the entry of new firms increases market supply and lowers the market price. But now each firm's entire ATC curve also shifts upward. The overall result is a higher-than-original equilibrium price. The industry produces a larger output at a higher product price because the industry expansion has increased resource prices and the minimum average total cost.

Since greater output will be supplied at a higher price, the long-run industry supply curve is upsloping. Instead of supplying 90,000, 100,000, or 110,000 units at the same price of $50, an increasing-cost industry might supply 90,000 units at $45, 100,000 units at $50, and 110,000 units at $55. A higher price is required to induce more production because costs per unit of output increase as production rises.

Figure 11.4 nicely illustrates the situation. Original market demand is D_1 and industry price and output are P_1 ($50) and Q_1 (100,000 units), respectively, at equilibrium point Y_1. An increase in demand to D_2 upsets this equilibrium and leads to economic profits. New firms enter the industry, increasing both market supply and the production costs of individual firms. A new price is established at point Y_2, where P_2 is $55 and Q_2 is 110,000 units.

Conversely, a decline in demand from D_1 to D_3 makes production unprofitable and causes firms to leave the industry. The resulting decline in resource prices reduces the minimum average total cost of production for firms that stay. A new equilibrium price is established at some level below the original price, say, at point Y_3, where P_3 is $45 and Q_3 is 90,000 units. Connecting these

FIGURE 11.4 The long-run supply curve for an increasing-cost industry. In an increasing-cost industry, the entry of new firms in response to an increase in demand (D_3 to D_1 to D_2) will bid up resource prices and thereby increase unit costs. As a result, an increased industry output (Q_3 to Q_1 to Q_2) will be forthcoming only at higher prices ($45 < $50 < $55). The long-run industry supply curve (S) therefore slopes upward through points Y_3, Y_1, and Y_2.



three equilibrium positions, we derive the upsloping long-run supply curve S in Figure 11.4.

Long-Run Supply for a Decreasing-Cost Industry

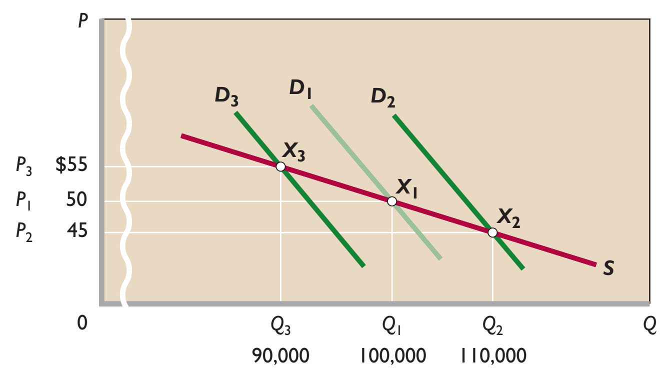
In **decreasing-cost industries,** firms experience lower costs as their industry expands. The personal computer industry is an example. As demand for personal computers increased, new manufacturers of computers entered the industry and greatly increased the resource demand for the components used to build them (for example, memory chips, hard drives, monitors, and operating software). The expanded production of the components enabled the producers of those items to achieve substantial economies of scale. The decreased production costs of the components reduced their prices, which greatly lowered the computer manufacturers' average costs of production. The supply of personal computers increased by more than demand, and the price of personal computers declined.

Unfortunately, however, the industries that show decreasing costs when output expands also show increasing costs if output contracts. A good example is the American shoe-manufacturing industry as it contracted due to foreign competition. Back when the industry was doing well and there were many shoemaking firms, the cost of specialized technicians who repair shoemaking machinery could be spread across many firms. This was because the repairmen worked as independent contractors going from one firm's factory to another firm's factory on a daily basis as various pieces of equipment at different factories needed repairs. But as the demand for American footwear fell over time, there were fewer and fewer factories, so the cost of a repair technician had to be spread over fewer and fewer firms. Thus, costs per firm and per unit of output increased.

Figure 11.5 illustrates the situation. The original market demand is D_1 and industry price and output are P_1 ($50) and Q_1 (100,000 units), respectively, at equilibrium point X_1. An increase in demand to D_2 upsets this equilibrium and leads to economic profits. New firms enter the industry, increasing market supply but decreasing the production costs of individual firms. A new price is established at point X_2, where P_2 is $45 and Q_2 is 110,000 units.

Conversely, a decline in demand from D_1 to D_3 makes production unprofitable and causes firms to leave the industry. The resulting increase in input prices increases the minimum average total cost of production for the firms that remain. A new equilibrium price is established at some level above the original price, say at point X_3, where P_3 is $55 and Q_3 is 90,000 units. Connecting these three equilibrium positions in Figure 11.5, we derive the downsloping long-run supply curve S for this decreasing-cost industry.

FIGURE 11.5 The long-run supply curve for a decreasing-cost industry. In a decreasing-cost industry, the entry of new firms in response to an increase in demand (D_3 to D_1 to D_2) will lead to decreased input prices and, consequently, decreased unit costs. As a result, an increase in industry output (Q_3 to Q_1 to Q_2) will be accompanied by lower prices ($55 > $50 > $45). The long-run industry supply curve (S) therefore slopes downward through points X_3, X_1, and X_2.



QUICK REVIEW 11.1

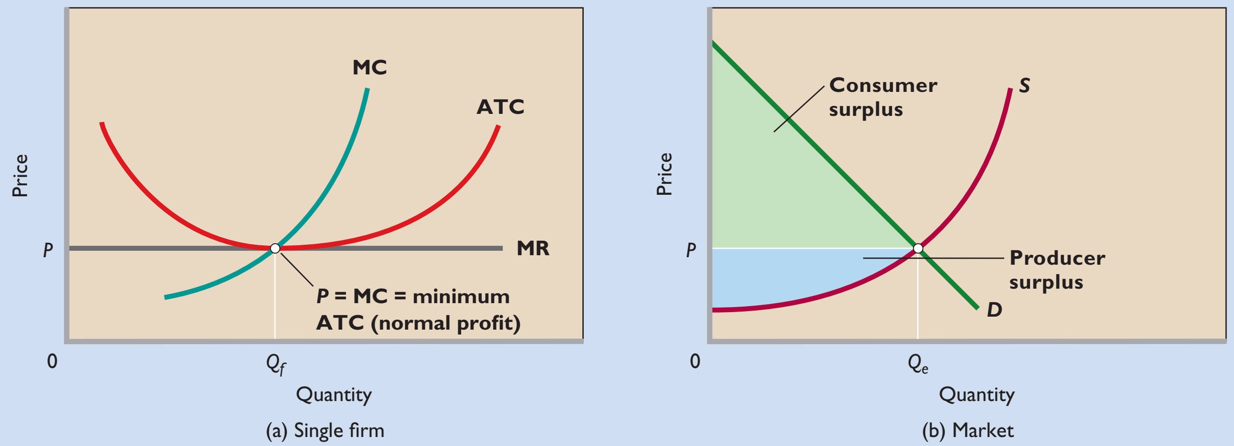
- In pure competition, entrepreneurs remove resources from industries and firms that are generating economic losses in order to transfer them to industries and firms that are generating economic profits.
- In the long run, the entry of firms into an industry will compete away any economic profits, and the exit of firms will eliminate economic losses, so price and minimum average total cost are equal. Entry and exit cease when the firms in the industry return to making a normal profit (zero economic profit).
- The long-run supply curves of constant-, increasing-, and decreasing-cost industries are horizontal, upsloping, and downsloping, respectively.

Pure Competition and Efficiency

LO11.4 Show how long-run equilibrium in pure competition produces an efficient allocation of resources.

Figure 11.6 (Key Graph) demonstrates the efficiency characteristics of the individual firms (Figure 11.6a) and the market (Figure 11.6b) after long-run adjustments in pure competition. Assuming a constant- or increasing-cost industry, the final long-run equilibrium positions of all firms have the same basic efficiency characteristics. As shown in

FIGURE 11.6 Long-run equilibrium: a competitive firm and market. (a) The equality of price (*P*), marginal cost (MC), and minimum average total cost (ATC) at output Q_f indicates that the firm is achieving productive efficiency and allocative efficiency. It is using the most efficient technology, charging the lowest price, and producing the greatest output consistent with its costs. It is receiving only a normal profit, which is incorporated into the ATC curve. The equality of price and marginal cost indicates that society allocated its scarce resources in accordance with consumer preferences. (b) In the purely competitive market, allocative efficiency occurs at the market equilibrium output Q_e. The sum of consumer surplus (green area) and producer surplus (blue area) is maximized.



QUICK QUIZ FOR FIGURE 11.6

1. We know the firm is a price taker because:
   - **a.** its MC curve slopes upward.
   - **b.** its ATC curve is U-shaped.
   - **c.** its MR curve is horizontal.
   - **d.** MC and ATC are equal at the profit-maximizing output.
2. At this firm's profit-maximizing output:
   - **a.** total revenue equals total cost.
   - **b.** it is earning an economic profit.
   - **c.** allocative, but not necessarily productive, efficiency is achieved.
   - **d.** productive, but not necessarily allocative, efficiency is achieved.
3. The equality of *P*, MC, and minimum ATC:
   - **a.** occurs only in constant-cost industries.
   - **b.** encourages entry of new firms.
   - **c.** means that the "right goods" are being produced in the "right ways."
   - **d.** results in a zero accounting profit.
4. When *P* = MC = lowest ATC for individual firms, in the market:
   - **a.** consumer surplus necessarily exceeds producer surplus.
   - **b.** consumer surplus plus producer surplus is at a maximum.
   - **c.** producer surplus necessarily exceeds consumer surplus.
   - **d.** supply and demand are identical.

Answers: 1. c; 2. a; 3. c; 4. b

Figure 11.6a, price (and marginal revenue) will settle where it is equal to minimum average total cost: *P* (and MR) = minimum ATC. Moreover, since the marginal-cost curve intersects the average-total-cost curve at its minimum point, marginal cost and average total cost are equal: MC = minimum ATC. So in long-run equilibrium a triple equality occurs: *P* (and MR) = MC = minimum ATC. Thus, in long-run equilibrium, each firm produces at the output level Q_f that is associated with this triple equality.[1]

The triple equality tells us two very important things about long-run equilibrium. First, it tells us that although a competitive firm may realize economic profit or loss in the short run, it will earn only a normal profit by producing in accordance with the MR (= *P*) = MC rule in the long run. Second, the triple equality tells us that in long-run equilibrium, the profit-maximizing decision rule that leads each firm to produce the quantity at which *P* = MR also implies that each firm will produce at the output level Q_f that is associated with the minimum point on each identical firm's ATC curve.

This is very important because it suggests that pure competition leads to the most efficient possible use of society's resources. Indeed, subject only to Chapter 4's

[1]This triple equality does not always hold for decreasing-cost industries in which individual firms produce a large fraction of the total market output. In such cases, MC may remain below ATC if average costs are decreasing. We will discuss this situation of "natural monopoly" in Chapter 12.

qualifications relating to public goods and externalities, an idealized purely competitive market economy composed of constant- or increasing-cost industries will generate both productive efficiency and allocative efficiency.

Productive Efficiency: P = Minimum ATC

Productive efficiency requires that goods be produced in the least costly way. In the long run, pure competition forces firms to produce at the minimum average total cost of production and to charge a price that is just consistent with that cost. This is true because firms that do not use the best available (least-cost) production methods and combinations of inputs will not survive.

To see why that is true, let's suppose that Figure 11.6 has to do with pure competition in the cucumber industry. In the final equilibrium position shown in Figure 11.6a, suppose each firm in the cucumber industry is producing 100 units (say, truckloads) of cucumbers by using \$5,000 (equal to average total cost of \$50 × 100 units) worth of resources. If any firm produced that same amount of output at any higher total cost, say \$7,000, it would be wasting resources because all of the other firms in the industry are able to produce that same amount of output using only \$5,000 of resources. Society would be faced with a net loss of \$2,000 worth of alternative products. But this cannot happen in pure competition; this firm would incur a loss of \$2,000, requiring it to either reduce its costs or go out of business.

Note, too, that consumers benefit from productive efficiency by paying the lowest product price possible under the prevailing technology and cost conditions. And the firm receives only a normal profit, which is part of its economic costs and thus incorporated in its ATC curve.

Allocative Efficiency: *P* = MC

Long-run equilibrium in pure competition guarantees productive efficiency, such that output will be produced in the least-cost way. But productive efficiency by itself does not guarantee that anyone will want to buy the items that are being produced in the least-cost manner. For all we know, consumers might prefer that the resources used to produce those items be redirected toward producing other products instead.

ORIGIN OF THE IDEA

O11.1

Allocative efficiency

Fortunately, long-run equilibrium in pure competition also guarantees **allocative efficiency,** so we can be certain that society's scarce resources are directed toward producing the goods and services that people most want to consume. Stated formally, allocative efficiency occurs when it is impossible to produce any net gains for society by altering the combination of goods and services that are produced from society's limited supply of resources.

To understand how pure competition leads to allocative efficiency, recall the concept of opportunity cost while looking at Figure 11.6b, where Q_e total units are being produced in equilibrium by the firms in a purely competitive industry. For every unit up to Q_e, market demand curve *D* lies above market supply curve *S*. Recall from Chapter 4 what this means in terms of marginal benefits and marginal costs.

- For each unit of output on the horizontal axis, the point directly above it on demand curve *D* shows how many dollars' worth of other goods and services consumers are willing to give up to obtain that unit of output. Consequently, the demand curve shows the dollar value of the marginal benefit that consumers place on each unit.
- For each unit of output on the horizontal axis, the point directly above it on supply curve *S* shows how many dollars' worth of other products have to be sacrificed in order to direct the underlying resources toward producing each unit of this product. Consequently, supply curve *S* shows the dollar value of the marginal opportunity cost of each unit.

Keeping these definitions in mind, the fact that the demand curve lies above the supply curve for every unit up to Q_e means that marginal benefit exceeds marginal cost for every one of these units. Stated slightly differently, producing and consuming these units brings net benefits because consumers are willing to give up more of other goods to obtain these units than must actually be forgone to produce them. Furthermore, because the supply curve includes the opportunity cost of the other goods that must be given up when resources are directed to producing these units, we can be certain that consumers prefer to have the necessary resources directed toward producing these units rather than anything else. In other words, allocative efficiency has been achieved because redirecting the necessary resources toward producing anything else would make people less happy.

The fact that pure competition yields allocative efficiency can also be understood by looking at the situation facing each individual firm in long-run equilibrium. To see this, take the market equilibrium price *P* that is determined in Figure 11.6b and see how it affects the behavior of the individual firm shown in Figure 11.6a. This profit-maximizing firm takes *P* as fixed and produces Q_f units, the output level at which *P* = MC.

By comparing the horizontal line at *P* with the upsloping MC curve, it is clear that for every unit up to Q_f, the price at which each unit can be sold exceeds the marginal cost of producing it. That is equivalent to saying that these units are worth more to consumers than they cost to make. Why? Because consumers are willing to forgo *P* dollars' worth of other goods and services when they pay *P* dollars for these units, but at the same time the firm only has to use less than *P* dollars' worth of resources to produce them. Thus, if these units are produced and consumed, there are net benefits and society comes out ahead. And, as with our previous analysis, allocative efficiency also obtains because by spending their *P* dollars per unit on these units rather than anything else, consumers are indicating that they would rather have the necessary resources directed toward producing these units rather than anything else.

Maximum Consumer and Producer Surplus

We confirm the existence of allocative efficiency in Figure 11.6b, where we see that pure competition maximizes the sum of the "benefit surpluses" to consumers and producers. Recall from Chapter 4 that **consumer surplus** is the difference between the maximum prices that consumers are willing to pay for a product (as shown by the demand curve) and the market price of the product. In Figure 11.6b, consumer surplus is the green triangle, which is the sum of the vertical distances between the demand curve and equilibrium price. In contrast, **producer surplus** is the difference between the minimum prices that producers are willing to accept for a product (as shown by the supply curve) and the market price of the product. Producer surplus is the sum of the vertical distances between the equilibrium price and the supply curve. Here producer surplus is the blue area.

At the equilibrium quantity Q_e, the combined amount of consumer surplus and producer surplus is maximized. Allocative efficiency occurs because, at Q_e, marginal benefit, reflected by points on the demand curve, equals marginal cost, reflected by points on the supply curve. Alternatively, the maximum willingness of consumers to pay for unit Q_e equals the minimum acceptable price of that unit to producers. At any output less than Q_e, the sum of consumer and producer surplus—the combined size of the green and blue area—would be less than that shown. At any output greater than Q_e, an efficiency loss (deadweight loss) would subtract from the combined consumer and producer surplus shown by the green and blue area.

After long-run adjustments, pure competition produces both productive and allocative efficiency. It yields a level of output at which $P = \text{MC} = \text{lowest ATC}$, marginal benefit = marginal cost, maximum willingness to pay for the last unit = minimum acceptable price for that unit, and combined consumer and producer surplus are maximized.

Dynamic Adjustments

A further attribute of purely competitive markets is their ability to restore the efficiency just described when disrupted by changes in the economy. A change in consumer tastes, resource supplies, or technology will automatically set in motion the appropriate realignments of resources. For example, suppose that cucumbers and pickles become dramatically more popular. First, the demand for cucumbers will increase in the market, increasing the price of cucumbers. So, at current output, the price of cucumbers will exceed their marginal cost. At this point efficiency will be lost, but the higher price will create economic profits in the cucumber industry and stimulate its expansion. The profitability of cucumbers will permit the industry to bid resources away from now-less-pressing uses, say, watermelons. Expansion of the industry will end only when the supply of cucumbers has expanded such that the price of cucumbers and their marginal cost are equal—that is, when allocative efficiency has been restored.

Similarly, a change in the supply of a particular resource—for example, the field laborers who pick cucumbers—or in a production technique will upset an existing price–marginal-cost equality by either raising or lowering marginal cost. The resulting inequality of MC and *P* will cause producers, in either pursuing profit or avoiding loss, to reallocate resources until product supply is such that price once again equals marginal cost. In so doing, they will correct any inefficiency in the allocation of resources that the original change may have temporarily imposed on the economy.

"Invisible Hand" Revisited

The highly efficient allocation of resources that a purely competitive economy promotes comes about because businesses and resource suppliers seek to further their self-interest. For private goods with no externalities (Chapter 4), the "invisible hand" (Chapter 2) is at work. The competitive system not only maximizes profits for individual producers but also, at the same time, creates a pattern of resource allocation that maximizes consumer satisfaction. The invisible hand thus organizes the private interests of producers in a way that is fully in sync with society's interest in using scarce resources efficiently. Striving to obtain a profit produces highly desirable economic outcomes.

LAST WORD

A Patent Failure?

Patents May Hinder Creative Destruction. If So, Should We Consider Abolishing Patents?

Patents give inventors the sole legal right to market and sell their new ideas for a period of 20 years. So when considering the pluses and minuses of the patent system, it is important to begin with the fact that the possibility of obtaining a patent gives inventors a strong financial incentive to bear the research and development (R&D) costs necessary to come up with innovative solutions to old problems.

At the same time, however, the patent system also gives patent holders the ability to stifle the creative energies of other inventors by suing or threating to sue any individual or firm that they believe is "infringing" on their patent by producing or utilizing their invention without permission.

The problem is most acute for products like cell phones that incorporate thousands of different technologies into a single product. That's because each of those technologies might possibly infringe on one or more patents. If so, a single lawsuit filed over just one of those patents could halt the production and sale of the entire product. The alleged infringement may be totally unintentional or a matter of honest dispute. But if a patent holder believes that some part of the phone is infringing on his patent, he can threaten to sue the manufacturer and demand the shutdown of all production unless he receives royalty payments in compensation.

Consider Microsoft, which 30 years ago was a successful innovator thanks to its Windows operating system. Over the last 10 years, however, its Windows-based cell phones have been a failure. Yet Microsoft CEO Steve Balmer threatened to shut down the production of all Android phones because the Android software used to run those extremely popular phones happens to incorporate the ability to schedule a meeting. That is a feature that most Android users don't even know about. But it is a functionality over which Microsoft holds a patent for mobile devices. So to avoid a lawsuit that could have shut down the production of all Android phones, Android's parent company, Google, is now paying Microsoft a licensing fee on each and every Android phone.

That situation is very problematic for creative destruction because the patent system is being used to help an old company that hasn't had a successful product in many years to effectively tax and benefit from the successful innovations of a young rival. That ability to tax is a form of life support that allows stodgy old firms to survive longer than they should against innovative rivals and the pressures of creative destruction.

Even worse, companies known as "patent trolls" have been created to buy up patents simply for the chance to sue other companies and collect royalties. The patent trolls invent nothing and produce nothing. But they are free under the current system to make billions of dollars every year by suing innovative companies.

In response, some economists have begun to argue that the net benefits of the patent system have been overstated and that innovation might proceed faster in in certain industries if patents were abolished. Their key insight is that the net benefits of

Technological Advance and Competition

LO11.5 Discuss creative destruction and the profit incentives for innovation.

In explaining the model of pure competition, we assumed for simplicity that all the firms in an industry had the same cost curves. Competition, as a result, only involved entrepreneurs entering and exiting industries in response to changes in profits caused by changes in the market price. This form of competition is important, but it is just a game of copycat because firms entering an industry simply duplicate the production methods and cost curves of existing firms in order to duplicate their above-normal profits. In this type of competition, there is no dynamism and no innovation, just more of the same.

By contrast, the most dynamic and interesting parts of competition are the fights between firms over the creation of new production technologies and new products. As we explain in detail in Web Chapter 13, firms have a strong profit incentive to develop both improved ways of making existing products as well as totally new products. To put that incentive in context, recall one fact that you just learned about long-run equilibrium in perfect competition. When each firm in a purely competitive industry has the same productive technology and therefore the same cost structure for producing output, entry and exit assure that in the long run every firm will make the exact same normal profit.

patents depend upon how easy it is for rivals to successfully copy and market an innovative product.

Consider pharmaceuticals. Once the chemical formula for a new drug becomes known, it is very easy for rivals to make chemically identical versions that will be easy to market because they will be just as effective as the version sold by the firm that invented the drug. At the same time, competition is so fierce in the pharmaceutical industry that without patent protection the price of the new drug would be driven down almost immediately to its marginal production cost. That is highly problematic because the market price would be too low to ever recoup the large R&D costs necessary to identify and develop effective new medications. Thus, without patent protection, R&D would cease and no new drugs would be developed.

So for industries like pharmaceuticals that have easy-to-copy products, patents should continue to exist as they are the only way to provide the financial incentive necessary to get firms to invest the R&D monies that must be spent if you want innovation and creative destruction.

Things are very different, however, for complicated consumer products that are made up of thousands of separate technologies that are each difficult to copy and market. As an example, even if Apple's rivals obtained the blueprints for the iPhone, it would still be extremely costly for them to build the factories necessary to make copies. And even if they did that, they would still have to convince consumers that their copycat iPhones were as good as the original. Thus, unlike pharmaceuticals, patents are not necessary to provide the firms that produce complicated consumer goods with an incentive to develop new products and invest in R&D.

On the other hand, society would likely see great benefits if patents were eliminated for complicated consumer goods like cell phones and automobiles because creative destruction would

likely increase as innovative companies would no longer fear patent-infringement lawsuits and old rivals could no longer delay their own demise by taxing innovators.

As a result, some economists now argue that patents should only be available for industries with simple products that are easy to copy and market. For industries with complicated products that are hard to copy and market, patents should be eliminated.

Entrepreneurs, of course, would like to earn more than a normal profit. As a result, they are constantly attempting two different strategies for increasing their profits. The first involves attempting to lower the production costs of existing products through better technology or improved business organization. Because pure competition implies that individual firms cannot affect the market price, anything that lowers an innovating firm's production costs will result in higher profits, since the innovating firm's revenues per unit (which are equal to the market price per unit) will stay the same while its costs per unit fall due to its improved production technology.

The second strategy for earning a rate of return greater than a normal profit is to try to develop a totally new product that is popular with consumers. If a firm is first-to-market with a popular new product, it will face no competition, as it is the only producer. As long as the product remains popular and the firm remains the only producer, it will be able to charge prices that are higher than production costs, thereby allowing it to earn above-normal profits. (We say much more about this in the next chapter, which covers pure monopoly).

Notably, however, any advantages that innovative firms gain either by lowering the production costs of existing products or by introducing entirely new products will not normally persist. An innovative entrepreneur may put some of her current rivals out of business, but there are always other entrepreneurs with new ideas so that soon it may be *her* firm that is going out of business due to innovations

CONSIDER THIS . . .

Running a Company Is Hard Business

The life expectancy of a U.S. business is just 10.2 years. About 9.5 percent of U.S. firms go out of business each year. In addition, 22 percent of new start-up firms go bankrupt within 2 years, 53 percent within 5 years, and nearly 65 percent within 10 years.

These numbers testify to the ability of competition to quickly dispose of firms that have high production costs or unpopular products. In a competitive environment, such firms quickly prove unprofitable and are shut down by their owners.

Balancing out the dying firms are start-ups that hope to use the resources freed up by the closed firms to deliver better products or lower costs. In a typical year, more than 650,000 new businesses are started in the United States. Most of these new firms will themselves eventually fall victim to creative destruction and the pressures of competition, but one of them may just be the next Google, Starbucks, or Walmart.

made by others. The nearby Consider This box shows just how rapidly new firms are created and destroyed.

Creative Destruction

The innovations that firms achieve thanks to competition are considered by many economists to be the driving force behind economic growth and rising living standards. The transformative effects of competition are often referred to as **creative destruction** to capture the idea that the creation of new products and new production methods destroys the market positions of firms committed to existing products and old ways of doing business. In addition, just the *threat* that a rival may soon come out with a new technology or product can cause other firms to innovate and thereby replace or rectify their old ways of doing business. As argued decades ago by Harvard economist Joseph Schumpeter, the most important type of competition is

ORIGIN OF THE IDEA

O11.2

Creative destruction

> competition from the new commodity, the new technology, the new source of supply, the new type of business organization—competition which commands a decisive cost or quality advantage and which strikes not at the margins of profits of the existing firms but at their foundation and their very lives. This kind of competition is . . . so . . . important that it becomes a matter of comparative indifference whether competition in the ordinary [short-run or long-run] sense functions more or less promptly. . . .
>
> . . . Competition of the kind we now have in mind acts not only when in being but also when it is merely an ever-present threat. It disciplines before it attacks. The businessman feels himself to be in a competitive situation even if he is alone in his field.[2]

There are many examples of creative destruction. In the 1800s wagons, ships, and barges were the only means of transporting freight until the railroads broke up their monopoly; the dominant market position of the railroads was, in turn, undermined by trucks and, later, by airplanes. Movies brought new competition to live theater, at one time the "only show in town." But movies were later challenged by broadcast television, which was then challenged by cable TV. Both are now challenged by Netflix, Amazon Instant Video, and other online video-on-demand services. Cassettes replaced records before compact discs undermined cassettes. Now iPods, MP3 players, and Internet music downloads will soon make the compact disc obsolete. Electronic communications—including faxes and e-mails—have pushed the U.S. Postal Service toward bankruptcy, including a $15.9 billion loss in 2012. And online retailers like Amazon.com have stolen substantial business away from brick-and-mortar retailers.

The "creative" part of "creative destruction" leads to new products and lower-cost production methods that are of great benefit to society because they allow for a more efficient use of society's scarce resources. Keep in mind, however, that the "destruction" part of "creative destruction" can be hard on workers in the industries being displaced by new technologies. A worker at a CD-making factory may see her job eliminated as consumers switch to online music downloads. The U.S. Postal Service cut over 160,000 jobs between 2008 and 2012 partly because of the impact that e-mail has had on the demand for postal services. And many jobs in retail have been eliminated due to competition with Amazon.com and other online retailers.

Normally, the process of creative destruction goes slowly enough that workers at firms being downsized can transition smoothly to jobs in firms that are expanding.

[2]Joseph A. Schumpeter, *Capitalism, Socialism, and Democracy*, 3d ed. (New York: Harper & Row, 1950), pp. 84–85.

But sometimes the change is too swift for all of them to find new jobs easily. And in other instances, such as a town with only one major employer—like a rural coal-mining town or a small town with a large auto factory—the loss of that one major employer can be devastating because there are not enough other firms in the local area to employ the workers laid off by the major employer.

While the net effects of creative destruction are indisputably positive—including ongoing economic growth and rising living standards—creative destruction involves costs as well as benefits. And while the benefits are widespread, the costs tend to be borne almost entirely by the relatively few workers in declining industries who are not positioned to make easy transitions to new jobs.

SUMMARY

LO11.1 Explain how the long run differs from the short run in pure competition.

In the short run, when plant and equipment are fixed, the firms in a purely competitive industry may earn profits or suffer losses. In the long run, when plant and equipment are adjustable, profits will attract new entrants, while losses will cause existing firms to leave the industry.

LO11.2 Describe how profits and losses drive the long-run adjustment process of pure competition.

The entry or exit of firms will change industry supply. Entry or exit will continue until the market price determined by industry supply interacting with market demand generates a normal profit for firms in the industry. With firms earning a normal profit, there will be no incentive to either enter or exit the industry. This situation constitutes long-run equilibrium in a purely competitive industry.

Entry and exit help to improve resource allocation. Firms that exit an industry due to low profits release their resources to be used more profitably in other industries. Firms that enter an industry chasing higher profits bring with them resources that were less profitably used in other industries. Both processes increase allocative efficiency.

In the long run, the market price of a product will equal the minimum average total cost of production. At a higher price, economic profits will cause firms to enter the industry until those profits have been competed away. At a lower price, losses will force the exit of firms from the industry until the product price rises to equal average total cost.

LO11.3 Explain the differences between constant-cost, increasing-cost, and decreasing-cost industries.

The long-run supply curve is horizontal for a constant-cost industry, upsloping for an increasing-cost industry, and downsloping for a decreasing-cost industry.

LO11.4 Show how long-run equilibrium in pure competition produces an efficient allocation of resources.

The long-run equality of price and minimum average total cost means that competitive firms will use the most efficient known technology and charge the lowest price consistent with their production costs. That is, the purely competitive firms will achieve productive efficiency.

The long-run equality of price and marginal cost implies that resources will be allocated in accordance with consumer tastes. Allocative efficiency will occur. In the market, the combined amount of consumer surplus and producer surplus will be at a maximum.

The competitive price system will reallocate resources in response to a change in consumer tastes, in technology, or in resource supplies and will thereby maintain allocative efficiency over time.

LO11.5 Discuss creative destruction and the profit incentives for innovation.

Competition involves never-ending attempts by entrepreneurs and managers to earn above-normal profits by either creating new products or developing lower-cost production methods for existing products. These efforts cause creative destruction, the financial undoing of the market positions of firms committed to existing products and old ways of doing business by new firms with new products and innovative ways of doing business.

TERMS AND CONCEPTS

long-run supply curve
constant-cost industry
increasing-cost industry
decreasing-cost industry
productive efficiency
allocative efficiency
consumer surplus
producer surplus
creative destruction

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Explain how the long run differs from the short run in pure competition. **LO11.1**
2. Relate opportunity costs to why profits encourage entry into purely competitive industries and how losses encourage exit from purely competitive industries. **LO11.2**
3. How do the entry and exit of firms in a purely competitive industry affect resource flows and long-run profits and losses? **LO11.2**
4. In long-run equilibrium, P = minimum ATC = MC. Of what significance for economic efficiency is the equality of P and minimum ATC? The equality of P and MC? Distinguish between productive efficiency and allocative efficiency in your answer. **LO11.4**
5. The basic model of pure competition reviewed in this chapter finds that in the long run all firms in a purely competitive industry will earn normal profits. If all firms will only earn a normal profit in the long run, why would any firms bother to develop new products or lower-cost production methods? Explain. **LO11.5**
6. "Ninety percent of new products fail within two years—so you shouldn't be so eager to innovate." Do you agree? Explain why or why not. **LO11.5**
7. **LAST WORD** How can patents speed up the process of creative destruction? How can patents slow down the process of creative destruction? How do differences in manufacturing costs affect which industries would be most likely to be affected by the removal of patents?

REVIEW QUESTIONS

1. When discussing pure competition, the term *long run* refers to a period of time long enough to allow: **LO11.1**
   - **a.** Firms already in an industry to either expand or contract their capacities.
   - **b.** New firms to enter or existing firms to leave.
   - **c.** Both *a* and *b*.
   - **d.** None of the above.
2. Suppose that the pen-making industry is perfectly competitive. Also suppose that each current firm and any potential firms that might enter the industry all have identical cost curves, with minimum ATC = \$1.25 per pen. If the market equilibrium price of pens is currently \$1.50, what would you expect it to be in the long run? **LO11.2**
   - **a.** \$0.25.
   - **b.** \$1.00.
   - **c.** \$1.25.
   - **d.** \$1.50.
3. Suppose that as the output of mobile phones increases, the cost of touch screens and other component parts decreases. If the mobile phone industry features pure competition, we would expect the long-run supply curve for mobile phones to be: **LO11.3**
   - **a.** Upward sloping.
   - **b.** Downward sloping.
   - **c.** Horizontal.
   - **d.** U-shaped.
4. Using diagrams for both the industry and a representative firm, illustrate competitive long-run equilibrium. Assuming constant costs, employ these diagrams to show how (*a*) an increase and (*b*) a decrease in market demand will upset that long-run equilibrium. Trace graphically and describe verbally the adjustment processes by which long-run equilibrium is restored. Now rework your analysis for increasing- and decreasing-cost industries and compare the three long-run supply curves. **LO11.3**
5. Suppose that purely competitive firms producing cashews discover that P exceeds MC. Is their combined output of cashews too little, too much, or just right to achieve allocative efficiency? In the long run, what will happen to the supply of cashews and the price of cashews? Use a supply and demand diagram to show how that response will change the combined amount of consumer surplus and producer surplus in the market for cashews. **LO11.4**

PROBLEMS

1. A firm in a purely competitive industry has a typical cost structure. The normal rate of profit in the economy is 5 percent. This firm is earning \$5.50 on every \$50 invested by its founders. What is its percentage rate of return? Is the firm earning an economic profit? If so, how large? Will this industry see entry or exit? What will be the rate of return earned by firms in this industry once the industry reaches long-run equilibrium? **LO11.2**
2. A firm in a purely competitive industry is currently producing 1,000 units per day at a total cost of \$450. If the firm produced 800 units per day, its total cost would be \$300, and if it produced 500 units per day, its total cost would be \$275. What are the firm's ATC per unit at these three levels of production? If every firm in this industry has the same cost structure, is the industry in long-run competitive equilibrium? From what you know about these firms' cost structures,

what is the highest possible price per unit that could exist as the market price in long-run equilibrium? If that price ends up being the market price and if the normal rate of profit is 10 percent, then how big will each firm's accounting profit per unit be? **LO11.4**

3. There are 300 purely competitive farms in the local dairy market. Of the 300 dairy farms, 298 have a cost structure that generates profits of $24 for every $300 invested. What is their percentage rate of return? The other two dairies have a cost structure that generates profits of $22 for every $200 invested. What is their percentage rate of return? Assuming that the normal rate of profit in the economy is 10 percent, will there be entry or exit? Will the change in the number of firms affect the two that earn $22 for every $200 invested? What will be the rate of return earned by most firms in the industry in long-run equilibrium? If firms can copy each other's technology, what will be the rate of return eventually earned by all firms? **LO11.4**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **12**

Pure Monopoly

Learning Objectives

LO12.1 List the characteristics of pure monopoly.

LO12.2 List and explain the barriers to entry that shield pure monopolies from competition.

LO12.3 Explain how demand is seen by a pure monopoly.

LO12.4 Explain how a pure monopoly sets its profit-maximizing output and price.

LO12.5 Discuss the economic effects of monopoly.

LO12.6 Describe why a monopolist might prefer to charge different prices in different markets.

LO12.7 Distinguish between the monopoly price, the socially optimal price, and the fair-return price of a government-regulated monopoly.

We turn now from pure competition to pure monopoly, which is at the opposite end of the spectrum of industry structures listed in Table 10.1. You deal with monopolies more often than you might think. If you see the logo for Microsoft's Windows on your computer, you are dealing with a monopoly (or, at least, a near-monopoly). When you purchase certain prescription drugs, you are buying monopolized products. When you make a local telephone call, turn on your lights, or subscribe to cable TV, you may be patronizing a monopoly, depending on your location.

What precisely do we mean by pure monopoly, and what conditions enable it to arise and survive? How does a pure monopolist determine its profit-maximizing price and output? Does a pure monopolist achieve the efficiency associated with pure competition? If not, what, if anything, should the government do about it? A simplified model of pure monopoly will help us answer these questions. It will be the first of three models of imperfect competition.

An Introduction to Pure Monopoly

LO12.1 List the characteristics of pure monopoly.

Pure monopoly exists when a single firm is the sole producer of a product for which there are no close substitutes. Here are the main characteristics of pure monopoly:

- ***Single seller*** A pure, or absolute, monopoly is an industry in which a single firm is the sole producer of a specific good or the sole supplier of a service; the firm and the industry are synonymous.
- ***No close substitutes*** A pure monopoly's product is unique in that there are no close substitutes. The consumer who chooses not to buy the monopolized product must do without it.
- ***Price maker*** The pure monopolist controls the total quantity supplied and thus has considerable control over price; it is a *price maker* (unlike a pure competitor, which has no such control and therefore is a *price taker*). The pure monopolist confronts the usual downsloping product demand curve. It can change its product price by changing the quantity of the product it produces. The monopolist will use this power whenever it is advantageous to do so.
- ***Blocked entry*** A pure monopolist has no immediate competitors because certain barriers keep potential competitors from entering the industry. Those barriers may be economic, technological, legal, or of some other type. But entry is totally blocked in pure monopoly.
- ***Nonprice competition*** The product produced by a pure monopolist may be either standardized (as with natural gas and electricity) or differentiated (as with Windows or Frisbees). Monopolists that have standardized products engage mainly in public relations advertising, whereas those with differentiated products sometimes advertise their products' attributes.

ORIGIN OF THE IDEA

O12.1

Monopoly

Examples of Monopoly

Examples of *pure* monopoly are relatively rare, but there are many examples of less pure forms. In most cities, government-owned or government-regulated public utilities—natural gas and electric companies, the water company, the cable TV company, and the local telephone company—are all monopolies or virtually so.

There are also many "near-monopolies" in which a single firm has the bulk of sales in a specific market. Intel, for example, produces 80 percent of the central microprocessors used in personal computers. First Data Corporation, via its Western Union subsidiary, accounts for 80 percent of the market for money order transfers. Brannock Device Company has an 80 percent market share of the shoe sizing devices found in shoe stores. Wham-O, through its Frisbee brand, sells 90 percent of plastic throwing disks. Google executes nearly 70 percent of all U.S. Internet searches and consequently controls nearly 75 percent of all the revenue generated by search ads in the United States (see this chapter's Last Word).

Professional sports teams are, in a sense, monopolies because they are the sole suppliers of specific services in large geographic areas. With a few exceptions, a single major-league team in each sport serves each large American city. If you want to see a live Major League Baseball game in St. Louis or Seattle, you must patronize the Cardinals or the Mariners, respectively. Other geographic monopolies exist. For example, a small town may be served by only one airline or railroad. In a small, isolated community, the local barber shop, dry cleaner, or grocery store may approximate a monopoly. And in the skies above, airlines control the only Internet access that is available to the passengers flying on on their planes.

Of course, there is almost always some competition. Satellite television is a substitute for cable, and amateur softball is a substitute for professional baseball. The Linux operating system can substitute for Windows, and so on. But such substitutes are typically either more costly or in some way less appealing.

Dual Objectives of the Study of Monopoly

Monopoly is worth studying both for its own sake and because it provides insights about the more common market structures of monopolistic competition and oligopoly (Chapter 13). These two market structures combine, in differing degrees, characteristics of pure competition and pure monopoly.

Barriers to Entry

LO12.2 List and explain the barriers to entry that shield pure monopolies from competition.

The factors that prohibit firms from entering an industry are called **barriers to entry.** In pure monopoly, strong barriers to entry effectively block all potential competition. Somewhat weaker barriers may permit oligopoly, a market structure dominated by a few firms. Still weaker barriers may permit the entry of a fairly large number of competing firms giving rise to monopolistic competition. And the absence of any effective entry barriers permits the entry of a very large number of firms, which provide the basis of pure competition. So barriers to entry are pertinent not only to the extreme case of pure monopoly but also to other market structures in which there are monopoly-like characteristics or monopoly-like behaviors.

We now discuss the four most prominent barriers to entry.

Economies of Scale

Modern technology in some industries is such that economies of scale—declining average total cost with added firm size—are extensive. In such cases, a firm's long-run average-cost schedule will decline over a wide range of output. Given market demand, only a few large firms or, in the extreme, only a single large firm can achieve low average total costs.

Figure 12.1 indicates economies of scale over a wide range of outputs. If total consumer demand is within that output range, then only a single producer can satisfy demand at least cost. Note, for example, that a monopolist can produce 200 units at a per-unit cost of \$10 and a total cost of \$2,000. If the industry has two firms and each produces 100 units, the unit cost is \$15 and total cost rises to \$3,000 (= 200 units × \$15). A still more competitive situation with four firms each producing 50 units would boost unit and total cost to \$20 and \$4,000, respectively. Conclusion: When long-run ATC is declining, only a single producer, a monopolist, can produce any particular amount of output at minimum total cost.

If a pure monopoly exists in such an industry, economies of scale will serve as an entry barrier and will protect the monopolist from competition. New firms that try to enter the industry as small-scale producers cannot realize the cost economies of the monopolist. They therefore will be undercut and forced out of business by the monopolist, which can sell at a much lower price and still make a profit because of its lower per-unit cost associated with its economies of scale. A new firm might try to start out big, that is, to enter the industry as a large-scale producer so as to achieve the necessary economies of scale. But the massive expense of the plant facilities along with customer loyalty to the existing product would make the entry highly risky. Therefore, the new and untried enterprise would find it difficult to secure financing for its venture. In most cases the risks and financial obstacles to "starting big" are prohibitive. This explains why efforts to enter such industries as computer operating software, commercial aircraft, and household laundry equipment are so rare.

A monopoly firm is referred to as a *natural monopoly* if the market demand curve intersects the long-run ATC curve at any point where average total costs are declining. If a natural monopoly were to set its price where market demand intersects long-run ATC, its price would be lower than if the industry were more competitive. But it will probably set a higher price. As with any monopolist, a natural monopolist may, instead, set its price far above ATC and obtain substantial economic profit. In that event, the lowest-unit-cost advantage of a natural monopolist

ORIGIN OF THE IDEA

O12.2

Natural monopoly

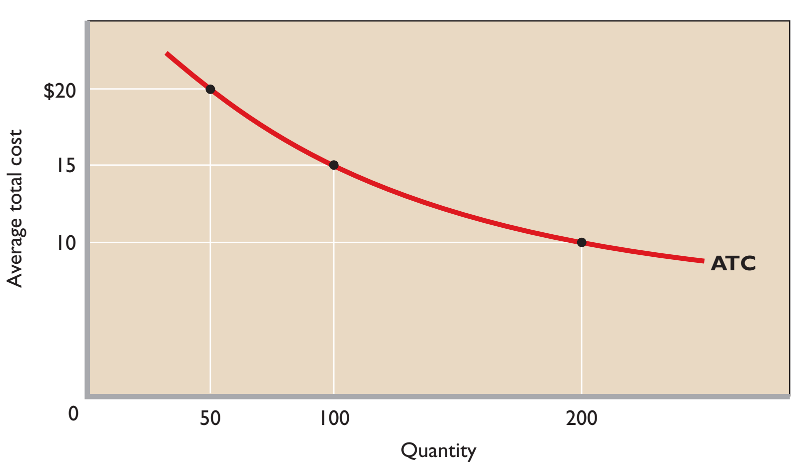


FIGURE 12.1 Economies of scale: the natural monopoly case. A declining long-run average-total-cost curve over a wide range of output quantities indicates extensive economies of scale. A single monopoly firm can produce, say, 200 units of output at lower cost (\$10 each) than could two or more firms that had a combined output of 200 units.

would accrue to the monopolist as profit and not as lower prices to consumers. That is why the government regulates some natural monopolies, specifying the price they may charge. We will say more about that later.

Legal Barriers to Entry: Patents and Licenses

Government also creates legal barriers to entry by awarding patents and licenses.

Patents A *patent* is the exclusive right of an inventor to use, or to allow another to use, her or his invention. Patents and patent laws aim to protect the inventor from rivals who would use the invention without having shared in the effort and expense of developing it. At the same time, patents provide the inventor with a monopoly position for the life of the patent. The world's nations have agreed on a uniform patent length of 20 years from the time of application. Patents have figured prominently in the growth of modern-day giants such as IBM, Pfizer, Intel, Xerox, General Electric, and DuPont.

Research and development (R&D) is what leads to most patentable inventions and products. Firms that gain monopoly power through their own research or by purchasing the patents of others can use patents to strengthen their market position. The profit from one patent can finance the research required to develop new patentable products. In the pharmaceutical industry, patents on prescription drugs have produced large monopoly profits that have helped finance the discovery of new patentable medicines. So monopoly power achieved through patents may well be self-sustaining, even though patents eventually expire and generic drugs then compete with the original brand. (Chapter 11's Last Word has more on the costs and benefits of patents.)

Licenses Government may also limit entry into an industry or occupation through *licensing*. At the national level, the Federal Communications Commission licenses only so many radio and television stations in each geographic area. In many large cities one of a limited number of municipal licenses is required to drive a taxicab. The consequent restriction of the supply of cabs creates economic profit for cab owners and drivers. New cabs cannot enter the industry to drive down prices and profits. In a few instances the government might "license" itself to provide some product and thereby create a public monopoly. For example, in some states only state-owned retail outlets can sell liquor. Similarly, many states have "licensed" themselves to run lotteries.

Ownership or Control of Essential Resources

A monopolist can use private property as an obstacle to potential rivals. For example, a firm that owns or controls a resource essential to the production process can prohibit the entry of rival firms. At one time the International Nickel Company of Canada (now called Vale Canada Limited) controlled 90 percent of the world's known nickel reserves. A local firm may own all the nearby deposits of sand and gravel. And it is very difficult for new sports leagues to be created because existing professional sports leagues have contracts with the best players and have long-term leases on the major stadiums and arenas.

Pricing and Other Strategic Barriers to Entry

Even if a firm is not protected from entry by, say, extensive economies of scale or ownership of essential resources, entry may effectively be blocked by the way the monopolist responds to attempts by rivals to enter the industry. Confronted with a new entrant, the monopolist may "create an entry barrier" by slashing its price, stepping up its advertising, or taking other strategic actions to make it difficult for the entrant to succeed.

Some examples of entry deterrence: In 2005 Dentsply, the dominant American maker of false teeth (80 percent market share) was found to have unlawfully precluded independent distributors of false teeth from carrying competing brands. The lack of access to the distributors deterred potential foreign competitors from entering the U.S. market. As another example, in 2001 a U.S. court of appeals upheld a lower court's finding that Microsoft used a series of illegal actions to maintain its monopoly in Intel-compatible PC operating systems (95 percent market share). One such action was charging higher prices for its Windows operating system to computer manufacturers that featured Netscape's Navigator Web browser rather than Microsoft's Internet Explorer.

Monopoly Demand

LO12.3 Explain how demand is seen by a pure monopoly.

Now that we have explained the sources of monopoly, we want to build a model of pure monopoly so that we can analyze its price and output decisions. Let's start by making three assumptions:

- Patents, economies of scale, or resource ownership secures the firm's monopoly.

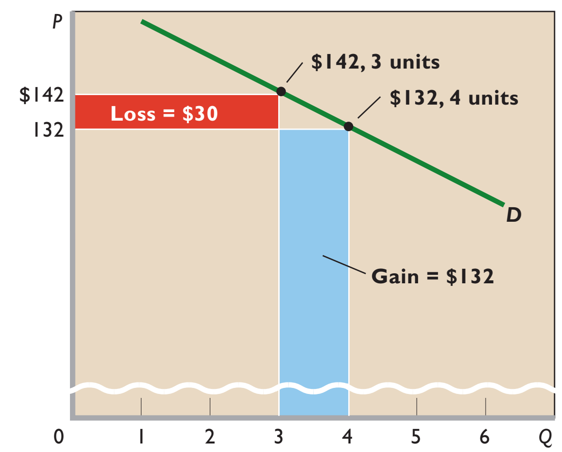
- No unit of government regulates the firm.
- The firm is a single-price monopolist; it charges the same price for all units of output.

The crucial difference between a pure monopolist and a purely competitive seller lies on the demand side of the market. The purely competitive seller faces a perfectly elastic demand at the price determined by market supply and demand. It is a price taker that can sell as much or as little as it wants at the going market price. Each additional unit sold will add the amount of the constant product price to the firm's total revenue. That means that marginal revenue for the competitive seller is constant and equal to product price. (Refer to the table and graph in Figure 10.1 for price, marginal-revenue, and total-revenue relationships for the purely competitive firm.)

The demand curve for the monopolist (and for any imperfectly competitive seller) is quite different from that of the pure competitor. Because the pure monopolist *is* the industry, its demand curve *is* the market demand curve. And because market demand is not perfectly elastic, the monopolist's demand curve is downsloping. Columns 1 and 2 in Table 12.1 illustrate this concept. Note that quantity demanded increases as price decreases.

In Figure 10.7 we drew separate demand curves for the purely competitive industry and for a single firm in such an industry. But only a single demand curve is needed in pure monopoly because the firm and the industry are one and the same. We have graphed part of the demand data in Table 12.1 as demand curve *D* in Figure 12.2. This is the monopolist's demand curve *and* the market demand curve. The downsloping demand curve has three implications that are essential to understanding the monopoly model.

FIGURE 12.2 Price and marginal revenue in pure monopoly. A pure monopolist, or any other imperfect competitor with a downsloping demand curve such as *D*, must set a lower price in order to sell more output. Here, by charging $132 rather than $142, the monopolist sells an extra unit (the fourth unit) and gains $132 from that sale. But from this gain must be subtracted $30, which reflects the $10 less the monopolist charged for each of the first 3 units. Thus, the marginal revenue of the fourth unit is $102 (= $132 − $30), considerably less than its $132 price.



Marginal Revenue Is Less Than Price

With a fixed downsloping demand curve, the pure monopolist can increase sales only by charging a lower price.

TABLE 12.1 Revenue and Cost Data of a Pure Monopolist

Revenue Data				Cost Data			
(1) Quantity of Output	(2) Price (Average Revenue)	(3) Total Revenue, (1) × (2)	(4) Marginal Revenue	(5) Average Total Cost	(6) Total Cost, (1) × (5)	(7) Marginal Cost	(8) Profit [+] or Loss [−]
0	$172	$ 0			$ 100		$−100
			$162			$ 90	
1	162	162		$190.00	190		−28
			142			80	
2	152	304		135.00	270		+34
			122			70	
3	142	426		113.33	340		+86
			102			60	
4	132	528		100.00	400		+128
			82			70	
5	122	610		94.00	470		+140
			62			80	
6	112	672		91.67	550		+122
			42			90	
7	102	714		91.43	640		+74
			22			110	
8	92	736		93.75	750		−14
			2			130	
9	82	738		97.78	880		−142
			−18			150	
10	72	720		103.00	1030		−310

Consequently, marginal revenue—the change in total revenue associated with a one-unit change in output—is less than price (average revenue) for every unit of output except the first. Why so? The reason is that the lower price of the extra unit of output also applies to all prior units of output. The monopolist could have sold these prior units at a higher price if it had not produced and sold the extra output. Each additional unit of output sold increases total revenue by an amount equal to its own price less the sum of the price cuts that apply to all prior units of output.

Figure 12.2 confirms this point. There, we have highlighted two price-quantity combinations from the monopolist's demand curve. The monopolist can sell 1 more unit at $132 than it can at $142 and that way obtain $132 (the blue area) of extra revenue. But to sell that fourth unit for $132, the monopolist must also sell the first 3 units at $132 rather than $142. The $10 reduction in revenue on 3 units results in a $30 revenue loss (the red area). Thus, the net difference in total revenue from selling a fourth unit is $102: the $132 gain from the fourth unit minus the $30 forgone on the first 3 units. This net gain (marginal revenue) of $102 from the fourth unit is clearly less than the $132 price of the fourth unit.

Column 4 in Table 12.1 shows that marginal revenue is always less than the corresponding product price in column 2, except for the first unit of output. Because marginal revenue is the change in total revenue associated with each additional unit of output, the declining amounts of marginal revenue in column 4 mean that total revenue increases at a diminishing rate (as shown in column 3).

We show the relationship between the monopolist's marginal-revenue curve and total-revenue curve in Figure 12.3. For this figure, we extended the demand and revenue data of columns 1 through 4 in Table 12.1, assuming that each successive $10 price cut elicits 1 additional unit of sales. That is, the monopolist can sell 11 units at $62, 12 units at $52, and so on.

Note that the monopolist's *MR* curve lies below the demand curve, indicating that marginal revenue is less than price at every output quantity but the very first unit. Observe also the special relationship between total revenue (shown in the lower graph) and marginal revenue (shown in the top graph). Because marginal revenue is the change in total revenue, marginal revenue is positive while total revenue is increasing. When total revenue reaches its maximum, marginal revenue is zero. When total revenue is diminishing, marginal revenue is negative.

The Monopolist Is a Price Maker

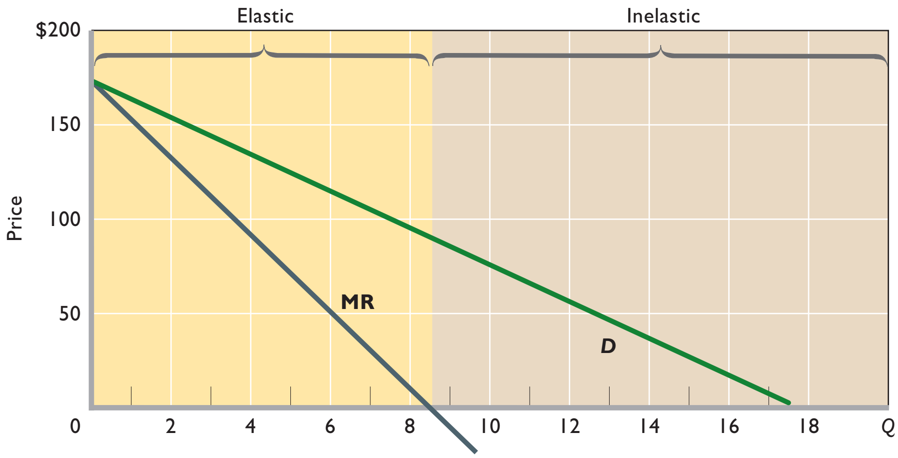
All imperfect competitors, whether pure monopolists, oligopolists, or monopolistic competitors, face downsloping demand curves. As a result, any change in quantity produced causes a movement along their respective demand curves and a change in the price they can charge for their respective products. Economists summarize this fact by saying that firms with downsloping demand curves are *price makers*.

This is most evident in pure monopoly, where an industry consists of a single monopoly firm so that total industry output is exactly equal to whatever the single monopoly firm chooses to produce. As we just mentioned, the monopolist faces a downsloping demand curve in which each amount of output is associated with some unique price. Thus, in deciding on the quantity of output to produce, the monopolist is also determining the price it will charge. Through control of output, it can "make the price." From columns 1 and 2 in Table 12.1 we find that the monopolist can charge a price of $72 if it produces and offers for sale 10 units, a price of $82 if it produces and offers for sale 9 units, and so forth.

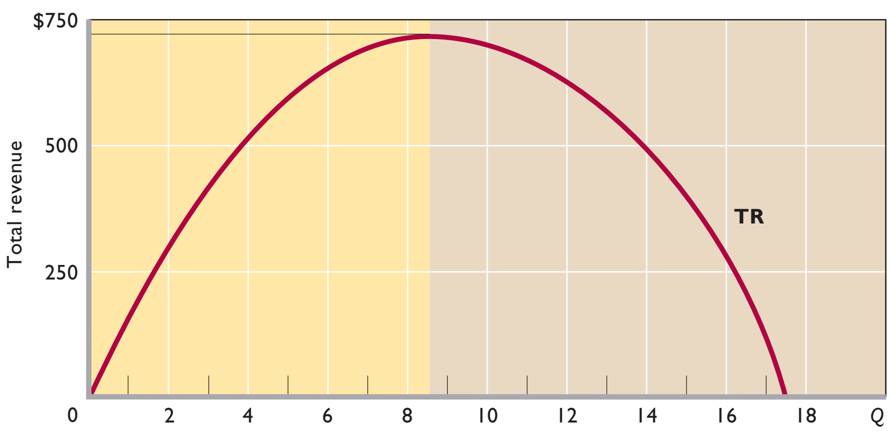
The Monopolist Sets Prices in the Elastic Region of Demand

The total-revenue test for price elasticity of demand is the basis for our third implication. Recall from Chapter 6 that the total-revenue test reveals that when demand is elastic, a decline in price will increase total revenue. Similarly, when demand is inelastic, a decline in price will reduce total revenue. Beginning at the top of demand curve *D* in Figure 12.3a, observe that as the price declines from $172 to approximately $82, total revenue increases (and marginal revenue therefore is positive). This means that demand is elastic in this price range. Conversely, for price declines below $82, total revenue decreases (marginal revenue is negative), indicating that demand is inelastic there.

The implication is that a monopolist will never choose a price-quantity combination where price reductions cause total revenue to decrease (marginal revenue to be negative). The profit-maximizing monopolist will always want to avoid the inelastic segment of its demand curve in favor of some price-quantity combination in the elastic region. Here's why: To get into the inelastic region, the monopolist must lower price and increase output. In the inelastic region a lower price means less total revenue. And increased output always means increased total cost. Less total revenue and higher total cost yield lower profit.



(a)
Demand and marginal-revenue curves



(b)
Total-revenue curve

FIGURE 12.3 Demand, marginal revenue, and total revenue for a pure monopolist. (a) Because it must lower price on all units sold in order to increase its sales, an imperfectly competitive firm's marginal-revenue curve (MR) lies below its downsloping demand curve (*D*). The elastic and inelastic regions of demand are highlighted. (b) Total revenue (TR) increases at a decreasing rate, reaches a maximum, and then declines. Note that in the elastic region, TR is increasing and hence MR is positive. When TR reaches its maximum, MR is zero. In the inelastic region of demand, TR is declining, so MR is negative.

QUICK REVIEW 12.1

- A pure monopolist is the sole supplier of a product or service for which there are no close substitutes.
- A monopoly survives because of entry barriers such as economies of scale, patents and licenses, the ownership of essential resources, and strategic actions to exclude rivals.
- The monopolist's demand curve is downsloping and its marginal-revenue curve lies below its demand curve.
- The downsloping demand curve means that the monopolist is a price maker.
- The monopolist will operate in the elastic region of demand since in the inelastic region it can increase total revenue and reduce total cost by reducing output.

Output and Price Determination

LO12.4 Explain how a pure monopoly sets its profit-maximizing output and price.

At what specific price-quantity combination will a profit-maximizing monopolist choose to operate? To answer this question, we must add production costs to our analysis.

Cost Data

On the cost side, we will assume that although the firm is a monopolist in the product market, it hires resources competitively and employs the same technology and, therefore, has the same cost structure as the purely competitive firm that we studied in Chapters 10 and 11. By using the same cost data that we developed in Chapter 9 and applied to the competitive firm in Chapters 10 and 11, we will be able to

KEY GRAPH

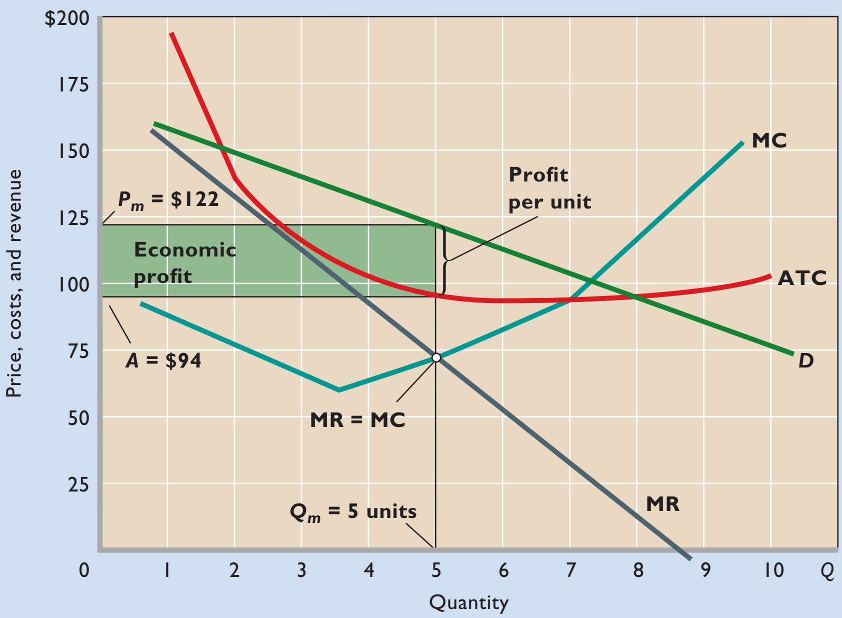


FIGURE 12.4 Profit maximization by a pure monopolist. The pure monopolist maximizes profit by producing at the MR = MC output, here $Q_m = 5$ units. Then, as seen from the demand curve, it will charge price $P_m = \$122$. Average total cost will be $A = \$94$, meaning that per-unit profit is $P_m - A$ and total profit is $5 \times (P_m - A)$. Total economic profit is thus represented by the green rectangle.

QUICK QUIZ FOR FIGURE 12.4

1. The MR curve lies below the demand curve in this figure because the:
   a. demand curve is linear (a straight line).
   b. demand curve is highly inelastic throughout its full length.
   c. demand curve is highly elastic throughout its full length.
   d. gain in revenue from an extra unit of output is less than the price charged for that unit of output.
2. The area labeled "Economic profit" can be found by multiplying the difference between *P* and ATC by quantity. It also can be found by:
   a. dividing profit per unit by quantity.
   b. subtracting total cost from total revenue.
   c. multiplying the coefficient of demand elasticity by quantity.
   d. multiplying the difference between *P* and MC by quantity.
3. This pure monopolist:
   a. charges the highest price that it could achieve.
   b. earns only a normal profit in the long run.
   c. restricts output to create an insurmountable entry barrier.
   d. restricts output to increase its price and total economic profit.
4. At this monopolist's profit-maximizing output:
   a. price equals marginal revenue.
   b. price equals marginal cost.
   c. price exceeds marginal cost.
   d. profit per unit is maximized.

Answers: 1. d; 2. b; 3. d; 4. c

directly compare the price and output decisions of a pure monopoly with those of a pure competitor. This will help us demonstrate that the price and output differences between a pure monopolist and a pure competitor are not the result of two different sets of costs. Columns 5 through 7 in Table 12.1 restate the pertinent cost data from Table 9.2.

MR = MC Rule

A monopolist seeking to maximize total profit will employ the same rationale as a profit-seeking firm in a competitive industry. If producing is preferable to shutting down, it will produce up to the output at which marginal revenue equals marginal cost (MR = MC).

A comparison of columns 4 and 7 in Table 12.1 indicates that the profit-maximizing output is 5 units because the fifth unit is the last unit of output whose marginal revenue exceeds its marginal cost. What price will the monopolist charge? The demand schedule shown as columns 1 and 2 in Table 12.1 indicates there is only one price at which 5 units can be sold: $122.

This analysis is shown in **Figure 12.4 (Key Graph),** where we have graphed the demand, marginal-revenue, average-total-cost, and marginal-cost data of Table 12.1. The profit-maximizing output occurs at 5 units of output (Q_m), where the marginal-revenue (MR) and marginal-cost (MC) curves intersect. There, MR = MC.

To find the price the monopolist will charge, we extend a vertical line from Q_m up to the demand curve D. The unique price P_m at which Q_m units can be sold is \$122. In this case, \$122 is the profit-maximizing price. So the monopolist sets the quantity at Q_m to charge its profit-maximizing price of \$122.

Columns 2 and 5 in Table 12.1 show that at 5 units of output, the product price (\$122) exceeds the average total cost (\$94). The monopolist thus obtains an economic profit of \$28 per unit, and the total economic profit is \$140 (= 5 units × \$28). In Figure 12.4, per-unit profit is $P_m - A$, where A is the average total cost of producing Q_m units. Total economic profit—the green rectangle—is found by multiplying this per-unit profit by the profit-maximizing output Q_m.

Another way to determine the profit-maximizing output is by comparing total revenue and total cost at each possible level of production and choosing the output with the greatest positive difference. Use columns 3 and 6 in Table 12.1 to verify that 5 units is the profit-maximizing output. An accurate graphing of total revenue and total cost against output would also show the greatest difference (the maximum profit) at 5 units of output. Table 12.2 summarizes the process for determining the profit-maximizing output, profit-maximizing price, and economic profit in pure monopoly.

WORKED PROBLEMS

W12.1

Monopoly price and output

No Monopoly Supply Curve

Recall that MR equals P in pure competition and that the supply curve of a purely competitive firm is determined by applying the MR (= P) = MC profit-maximizing rule. At any specific market-determined price, the purely competitive seller will maximize profit by supplying the quantity at which MC is equal to that price. When the market price increases or decreases, the competitive firm produces more or less output. Each market price is thus associated with a specific output, and all such price-output pairs define the supply curve. This supply curve turns out to be the portion of the firm's MC curve that lies above the average-variable-cost curve (see Figure 10.6).

At first glance we would suspect that the pure monopolist's marginal-cost curve would also be its supply curve. But that is *not* the case. *The pure monopolist has no supply curve.* There is no unique relationship between price and quantity supplied for a monopolist. Like the competitive firm, the monopolist equates marginal revenue and marginal cost to determine output, but for the monopolist marginal revenue is less than price. Because the monopolist does not equate marginal cost to price, it is possible for different demand conditions to bring about different prices for the same output. To understand this point, refer to Figure 12.4 and pencil in a new, steeper marginal-revenue curve that intersects the marginal-cost curve at the same point as does the present marginal-revenue curve. Then draw in a new demand curve that is roughly consistent with your new marginal-revenue curve. With the new curves, the same MR = MC output of 5 units now means a higher profit-maximizing price. Conclusion: There is no single, unique price associated with each output level Q_m, and so there is no supply curve for the pure monopolist.

Misconceptions Concerning Monopoly Pricing

Our analysis exposes two fallacies concerning monopoly behavior.

Not Highest Price Because a monopolist can manipulate output and price, people often believe it "will charge

TABLE 12.2 Steps for Graphically Determining the Profit-Maximizing Output, Profit-Maximizing Price, and Economic Profit (if Any) in Pure Monopoly

Step 1.	Determine the profit-maximizing output by finding where MR = MC.		
Step 2.	Determine the profit-maximizing price by extending a vertical line upward from the output determined in step 1 to the pure monopolist's demand curve.		
Step 3.	Determine the pure monopolist's economic profit using one of two methods:		
	Method 1.	Find profit per unit by subtracting the average total cost of the profit-maximizing output from the profit-maximizing price. Then multiply the difference by the profit-maximizing output to determine economic profit (if any).	
	Method 2.	Find total cost by multiplying the average total cost of the profit-maximizing output by that output. Find total revenue by multiplying the profit-maximizing output by the profit-maximizing price. Then subtract total cost from total revenue to determine economic profit (if any).	

the highest price possible." That is incorrect. There are many prices above P_m in Figure 12.4, but the monopolist shuns them because they yield a smaller-than-maximum total profit. The monopolist seeks maximum total profit, not maximum price. Some high prices that could be charged would reduce sales and total revenue too severely to offset any decrease in total cost.

Total, Not Unit, Profit The monopolist seeks maximum *total* profit, not maximum *unit* profit. In Figure 12.4 a careful comparison of the vertical distance between average total cost and price at various possible outputs indicates that per-unit profit is greater at a point slightly to the left of the profit-maximizing output Q_m. This is seen in Table 12.1, where the per-unit profit at 4 units of output is \$32 (= \$132 − \$100) compared with \$28 (= \$122 − \$94) at the profit-maximizing output of 5 units. Here the monopolist accepts a lower-than-maximum per-unit profit because additional sales more than compensate for the lower unit profit. A monopolist would rather sell 5 units at a profit of \$28 per unit (for a total profit of \$140) than 4 units at a profit of \$32 per unit (for a total profit of only \$128).

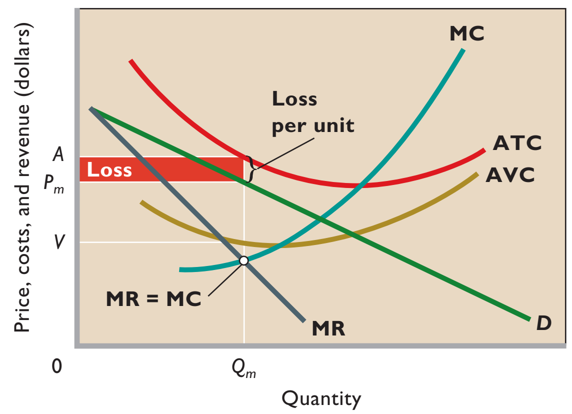
Possibility of Losses by Monopolist

The likelihood of economic profit is greater for a pure monopolist than for a pure competitor. In the long run the pure competitor is destined to have only a normal profit, whereas barriers to entry mean that any economic profit realized by the monopolist can persist. In pure monopoly there are no new entrants to increase supply, drive down price, and eliminate economic profit.

But pure monopoly does not guarantee profit. The monopolist is not immune from changes in tastes that reduce the demand for its product. Nor is it immune from upward-shifting cost curves caused by escalating resource prices. If the demand and cost situation faced by the monopolist is far less favorable than that in Figure 12.4, the monopolist will incur losses in the short run. Consider the monopoly enterprise shown in Figure 12.5. Despite its dominance in the market (as, say, a seller of home sewing machines), it suffers a loss, as shown, because of weak demand and relatively high costs. Yet it continues to operate for the time being because its total loss is less than its fixed cost. More precisely, at output Q_m the monopolist's price P_m exceeds its average variable cost V. Its loss per unit is $A - P_m$, and the total loss is shown by the red rectangle.

Like the pure competitor, the monopolist will not persist in operating at a loss. Faced with continuing losses, in the long run the firm's owners will move their resources to alternative industries that offer better profit opportunities.

FIGURE 12.5 The loss-minimizing position of a pure monopolist. If demand D is weak and costs are high, the pure monopolist may be unable to make a profit. Because P_m exceeds V, the average variable cost at the MR = MC output Q_m, the monopolist will minimize losses in the short run by producing at that output. The loss per unit is $A - P_m$, and the total loss is indicated by the red rectangle.



A monopolist such as the one depicted in Figure 12.5 must obtain a minimum of a normal profit in the long run or it will go out of business.

Economic Effects of Monopoly

LO12.5 Discuss the economic effects of monopoly.

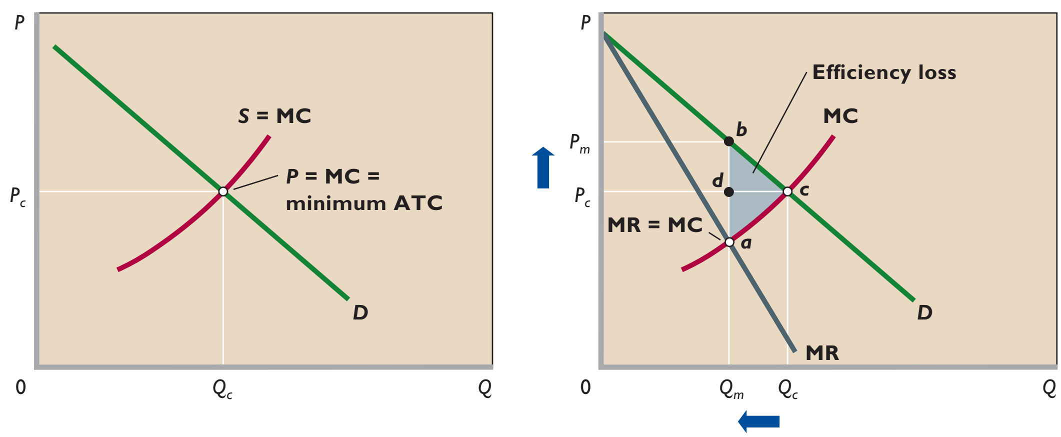
Let's now evaluate pure monopoly from the standpoint of society as a whole. Our reference for this evaluation will be the outcome of long-run efficiency in a purely competitive market, identified by the triple equality P = MC = minimum ATC.

Price, Output, and Efficiency

Figure 12.6 graphically contrasts the price, output, and efficiency outcomes of pure monopoly and a purely competitive *industry*. The S = MC curve in Figure 12.6a reminds us that the market supply curve S for a purely competitive industry is the horizontal sum of the marginal-cost curves of all the firms in the industry. Suppose there are 1,000 such firms. Comparing their combined supply curves S with market demand D, we see that the purely competitive price and output are P_c and Q_c.

Recall that this price-output combination results in both productive efficiency and allocative efficiency. *Productive efficiency* is achieved because free entry and exit force firms to operate where average total cost is at a minimum. The sum of the minimum-ATC outputs of the 1,000 pure competitors is the industry output, here, Q_c. Product price is at the lowest level consistent with minimum

FIGURE 12.6 Inefficiency of pure monopoly relative to a purely competitive industry. (a) In a purely competitive industry, entry and exit of firms ensure that price (P_c) equals marginal cost (MC) and that the minimum average-total-cost output (Q_c) is produced. Both productive efficiency (P = minimum ATC) and allocative efficiency (P = MC) are obtained. (b) In pure monopoly, the MR curve lies below the demand curve. The monopolist maximizes profit at output Q_m, where MR = MC, and charges price P_m. Thus, output is lower (Q_m rather than Q_c) and price is higher (P_m rather than P_c) than they would be in a purely competitive industry. Monopoly is inefficient, since output is less than that required for achieving minimum ATC (here, at Q_c) and because the monopolist's price exceeds MC. Monopoly creates an efficiency loss (here, of triangle *abc*). There is also a transfer of income from consumers to the monopoly (here, of rectangle P_cP_mbd).



(a)
Purely competitive industry

(b)
Pure monopoly

average total cost. The *allocative efficiency* of pure competition results because production occurs up to that output at which price (the measure of a product's value or marginal benefit to society) equals marginal cost (the worth of the alternative products forgone by society in producing any given commodity). In short: P = MC = minimum ATC.

Now let's suppose that this industry becomes a pure monopoly (Figure 12.6b) as a result of one firm acquiring all its competitors. We also assume that no changes in costs or market demand result from this dramatic change in the industry structure. What formerly were 1,000 competing firms is now a single pure monopolist consisting of 1,000 noncompeting branches.

The competitive market supply curve S has become the marginal-cost curve (MC) of the monopolist, the summation of the individual marginal-cost curves of its many branch plants. (Since the monopolist does not have a supply curve, as such, we have removed the S label.) The important change, however, is on the demand side. From the viewpoint of each of the 1,000 individual competitive firms, demand was perfectly elastic, and marginal revenue was therefore equal to the market equilibrium price P_c. So each firm equated its marginal revenue of P_c dollars per unit with its individual marginal cost curve to maximize profits. But market demand and individual demand are the same to the pure monopolist. The firm *is* the industry, and thus the monopolist sees the downsloping demand curve D shown in Figure 12.6b.

This means that marginal revenue is less than price, that graphically the MR curve lies below demand curve D. In using the MR = MC rule, the monopolist selects output Q_m and price P_m. A comparison of both graphs in Figure 12.6 reveals that the monopolist finds it profitable to sell a smaller output at a higher price than do the competitive producers.

Monopoly yields neither productive nor allocative efficiency. The lack of productive efficiency can be understood most directly by noting that the monopolist's output Q_m is less than Q_c, the output at which average total cost is lowest. In addition, the monopoly price P_m is higher than the competitive price P_c that we know in long-run equilibrium in pure competition equals minimum average total cost. Thus, the monopoly price exceeds minimum average total cost, thereby demonstrating in another way that the monopoly will not be productively efficient.

The monopolist's underproduction also implies allocative inefficiency. One way to see this is to note that at the monopoly output level Q_m, the monopoly price P_m that consumers are willing to pay exceeds the marginal cost of production. This means that consumers value additional units of this product more highly than they do the alternative products that could be produced from the resources that would be necessary to make more units of the monopolist's product.

The monopolist's allocative inefficiency can also be understood by noting that for every unit between Q_m and Q_c, marginal benefit exceeds marginal cost because the demand curve lies above the supply curve. By choosing not to produce these units, the monopolist reduces allocative efficiency because the resources that should have been used to make these units will be redirected instead toward producing items that bring lower net benefits to society. The total dollar value of this efficiency loss (or *deadweight loss*) is equal to the area of the gray triangle labeled *abc* in Figure 12.6b.

Income Transfer

In general, a monopoly transfers income from consumers to the owners of the monopoly. The income is received by the owners as revenue. Because a monopoly has market power, it can charge a higher price than would a purely competitive firm with the same costs. So the monopoly in effect levies a "private tax" on consumers. This private tax can often generate substantial economic profits that can persist because entry to the industry is blocked.

The transfer from consumers to the monopolist is evident in Figure 12.6b. For the Q_m units of output demanded, consumers pay price P_m rather than the price P_c that they would pay to a pure competitor. The total amount of income transferred from consumers to the monopolist is $P_m - P_c$ multiplied by the number of units sold, Q_m. So the total transfer is the dollar amount of rectangle P_cP_mbd. What the consumer loses, the monopolist gains. In contrast, the efficiency loss *abc* is a *deadweight* loss—society totally loses the net benefits of the Q_c minus Q_m units that are not produced.

Cost Complications

Our evaluation of pure monopoly has led us to conclude that, given identical costs, a purely monopolistic industry will charge a higher price, produce a smaller output, and allocate economic resources less efficiently than a purely competitive industry. These inferior results are rooted in the entry barriers characterizing monopoly.

Now we must recognize that costs may not be the same for purely competitive and monopolistic producers. The unit cost incurred by a monopolist may be either larger or smaller than that incurred by a purely competitive firm. There are four reasons why costs may differ: (1) economies of scale, (2) a factor called "X-inefficiency," (3) the need for monopoly-preserving expenditures, and (4) the "very long run" perspective, which allows for technological advance.

Economies of Scale Once Again Where economies of scale are extensive, market demand may not be sufficient to support a large number of competing firms, each producing at minimum efficient scale. In such cases, an industry of one or two firms would have a lower average total cost than would the same industry made up of numerous competitive firms. At the extreme, only a single firm—a natural monopoly—might be able to achieve the lowest long-run average total cost.

Some firms relating to new information technologies—for example, computer software, Internet service, and wireless communications—have displayed extensive economies of scale. As these firms have grown, their long-run average total costs have declined because of greater use of specialized inputs, the spreading of product development costs, and learning by doing. Also, *simultaneous consumption* and *network effects* have reduced costs.

A product's ability to satisfy a large number of consumers at the same time is called **simultaneous consumption** (or *nonrivalrous consumption*). Dell Computers needs to produce a personal computer for each customer, but Microsoft needs to produce its Windows program only once. Then, at very low marginal cost, Microsoft delivers its program by disk or Internet to millions of consumers. A similarly low cost of delivering product to additional customers is true for Internet service providers, music producers, and wireless communication firms. Because marginal costs are so low, the average total cost of output declines as more customers are added.

Network effects are present if the value of a product to each user, including existing users, increases as the total number of users rises. Good examples are computer software, cell phones, and Web sites like Facebook where the content is provided by users. When other people have Internet service and devices to access it, a person can conveniently send e-mail messages to them. And when they have similar software, various documents, spreadsheets, and photos can be attached to the e-mail messages. The greater the number of persons connected to the system, the more the benefits of the product to each person are magnified.

Such network effects may drive a market toward monopoly because consumers tend to choose standard products that everyone else is using. The focused demand for these products permits their producers to grow rapidly and thus achieve economies of scale. Smaller firms, which either have higher-cost "right" products or "wrong" products, get acquired or go out of business.

Economists generally agree that some new information firms have not yet exhausted their economies of scale. But most economists question whether such firms are truly natural monopolies. Most firms eventually achieve their minimum efficient scale at less than the full size of the market. That means competition among firms is possible.

But even if natural monopoly develops, the monopolist is unlikely to pass cost reductions along to consumers as price reductions. So, with perhaps a handful of exceptions, economies of scale do not change the general conclusion that monopoly industries are inefficient relative to competitive industries.

X-Inefficiency In constructing all the average-total-cost curves used in this book, we have assumed that the firm uses the most efficient existing technology. This assumption is only natural because firms cannot maximize profits unless they are minimizing costs. **X-inefficiency** occurs when a firm produces output at a higher cost than is necessary to produce it. In Figure 12.7 X-inefficiency is represented by operation at points X and X' above the lowest-cost ATC curve. At these points, per-unit costs are ATC_X (as opposed to ATC_1) for output Q_1 and $ATC_{X'}$ (as opposed to ATC_2) for output Q_2. Producing at any point above the average-total-cost curve in Figure 12.7 reflects inefficiency or "bad management" by the firm.

ORIGIN OF THE IDEA

O12.3

X-inefficiency

Why is X-inefficiency allowed to occur if it reduces profits? The answer is that managers may have goals, such as expanding power, an easier work life, avoiding business risk, or giving jobs to incompetent relatives, that conflict with cost minimization. Or X-inefficiency may arise because a firm's workers are poorly motivated or ineffectively supervised. Or a firm may simply become lethargic and inert, relying on rules of thumb in decision making as opposed to careful calculations of costs and revenues.

For our purposes the relevant question is whether monopolistic firms tend more toward X-inefficiency than competitive producers do. Presumably they do. Firms in competitive industries are continually under pressure from rivals, forcing them to be internally efficient to survive. But monopolists are sheltered from such competitive forces by entry barriers. That lack of pressure may lead to X-inefficiency.

Rent-Seeking Expenditures **Rent-seeking behavior** is any activity designed to transfer income or wealth to a particular firm or resource supplier at someone else's, or even society's, expense. We have seen that a monopolist can obtain an economic profit even in the long run. Therefore, it is no surprise that a firm may go to great expense to acquire or maintain a monopoly granted by government through legislation or an exclusive license. Such rent-seeking expenditures add nothing to the firm's output, but they clearly increase its costs. Taken alone, rent-seeking implies that monopoly involves even higher costs and even less efficiency than suggested in Figure 12.6b.

Technological Advance In the very long run, firms can reduce their costs through the discovery and implementation of new technology. If monopolists are more likely than competitive producers to develop more efficient production techniques over time, then the inefficiency of monopoly might be overstated. Because research and development (R&D) is the topic of optional Web Chapter 13, we will provide only a brief assessment here.

The general view of economists is that a pure monopolist will not be technologically progressive. Although its economic profit provides ample means to finance research

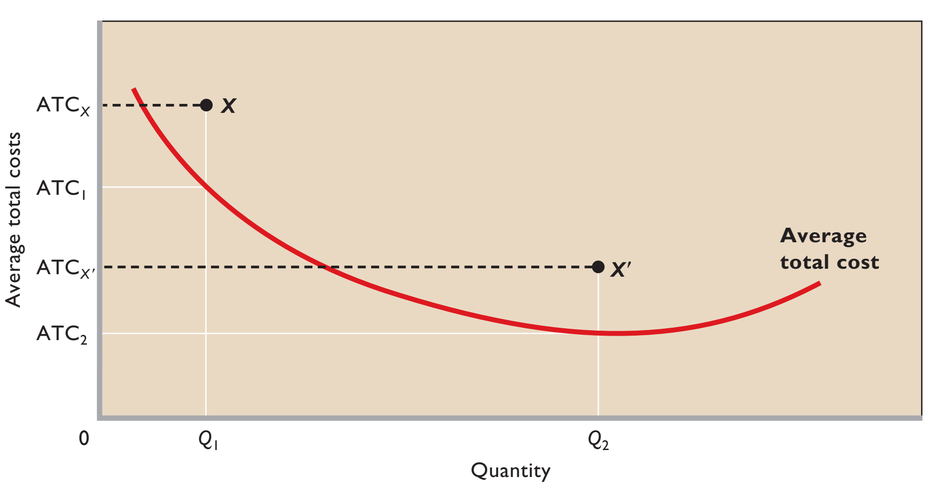


FIGURE 12.7 X-inefficiency. The average-total-cost curve (ATC) is assumed to reflect the minimum cost of producing each particular level of output. Any point above this "lowest-cost" ATC curve, such as X or X', implies X-inefficiency: operation at greater than lowest cost for a particular level of output.

and development, it has little incentive to implement new techniques (or products). The absence of competitors means that there is no external pressure for technological advance in a monopolized market. Because of its sheltered market position, the pure monopolist can afford to be complacent and lethargic. There simply is no major penalty for not being innovative.

One caveat: Research and technological advance may be one of the monopolist's barriers to entry. Thus, the monopolist may continue to seek technological advance to avoid falling prey to new rivals. In this case technological advance is essential to the maintenance of monopoly. But then it is *potential* competition, not the monopoly market structure, that is driving the technological advance. By assumption, no such competition exists in the pure monopoly model; entry is completely blocked.

GLOBAL PERSPECTIVE 12.1

Competition from Foreign Multinational Corporations

Competition from foreign multinational corporations diminishes the market power of firms in the United States. Here are just a few of the hundreds of foreign multinational corporations that compete strongly with U.S. firms in certain American markets.

Company (Country)	Main Products
Bayer (Germany)	chemicals
Daimler (Germany)	automobiles
Michelin (France)	tires
Lenovo (China)	electronics
Nestlé (Switzerland)	food products
Nokia (Finland)	wireless phones
Panasonic (Japan)	electronics
Petrobras (Brazil)	gasoline
Royal Dutch Shell (Netherlands)	gasoline
Samsung (South Korea)	electronics
Toyota (Japan)	automobiles

Source: Compiled from the Fortune 500 listing of the world's largest firms, "FORTUNE Global 500," **www.fortune.com**.

Assessment and Policy Options

Monopoly is a legitimate concern. Monopolists can charge higher-than-competitive prices that result in an underallocation of resources to the monopolized product. They can stifle innovation, engage in rent-seeking behavior, and foster X-inefficiency. Even when their costs are low because of economies of scale, there is no guarantee that the price they charge will reflect those low costs. The cost savings may simply accrue to the monopoly as greater economic profit.

Fortunately, however, monopoly is not widespread in the United States. Barriers to entry are seldom completely successful. Although research and technological advance may strengthen the market position of a monopoly, technology may also undermine monopoly power. Over time, the creation of new technologies may work to destroy monopoly positions. For example, the development of courier delivery, fax machines, and e-mail has eroded the monopoly power of the U.S. Postal Service. Similarly, cable television monopolies are now challenged by satellite TV and by technologies that permit the transmission of audio and video over the Internet.

Patents eventually expire; and even before they do, the development of new and distinct substitutable products often circumvents existing patent advantages. New sources of monopolized resources sometimes are found and competition from foreign firms may emerge. (See Global Perspective 12.1.) Finally, if a monopoly is sufficiently fearful of future competition from new products, it may keep its prices relatively low so as to discourage rivals from developing such products. If so, consumers may pay nearly competitive prices even though competition is currently lacking.

So what should government do about monopoly when it arises in the real world? Economists agree that government needs to look carefully at monopoly on a case-by-case basis. Three general policy options are available:

- If the monopoly is achieved and sustained through anticompetitive actions, creates substantial economic inefficiency, and appears to be long-lasting, the government can file charges against the monopoly under the antitrust laws. If found guilty of monopoly abuse, the firm can either be expressly prohibited from engaging in certain business activities or be broken into two or more competing firms. An example of the breakup approach was the dissolution of Standard Oil into several competing firms in 1911. In contrast, in 2001 an appeals court overruled a lower-court decision to divide Microsoft into two firms. Instead, Microsoft was prohibited from engaging in a number of specific anticompetitive business activities. (We discuss the antitrust laws and the Microsoft case in Chapter 19.)

- If the monopoly is a natural monopoly, society can allow it to continue to expand. If no competition emerges from new products, government may then decide to regulate its prices and operations. (We discuss this option later in this chapter and also in Chapter 19.)
- If the monopoly appears to be unsustainable because of emerging new technology, society can simply choose to ignore it. In such cases, society simply lets the process of creative destruction (discussed in Chapter 11) do its work. In Web Chapter 13, we discuss in detail the likelihood that real-world monopolies will collapse due to creative destruction and competition brought on by new technologies.

QUICK REVIEW 12.2

- The monopolist maximizes profit (or minimizes loss) at the output where MR = MC and charges the price that corresponds to that output on its demand curve.
- The monopolist has no supply curve, since any of several prices can be associated with a specific quantity of output supplied.
- Assuming identical costs, a monopolist will be less efficient than a purely competitive industry because it will fail to produce units of output for which marginal benefits exceed marginal costs.
- The inefficiencies of monopoly may be offset or lessened by economies of scale and, less likely, by technological progress, but they may be intensified by the presence of X-inefficiency and rent-seeking expenditures.

Price Discrimination

LO12.6 Describe why a monopolist might prefer to charge different prices in different markets.

We have assumed in this chapter that the monopolist charges a single price to all buyers. But under certain conditions the monopolist can increase its profit by charging different prices to different buyers. In so doing, the monopolist is engaging in **price discrimination,** the practice of selling a specific product at more than one price when the price differences are not justified by cost differences. Price discrimination can take three forms:

ORIGIN OF THE IDEA

O12.4

Price discrimination

- Charging each customer in a single market the maximum price she or he is willing to pay.
- Charging each customer one price for the first set of units purchased and a lower price for subsequent units purchased.
- Charging some customers one price and other customers another price.

Conditions

The opportunity to engage in price discrimination is not readily available to all sellers. Price discrimination is possible when the following conditions are met:

- ***Monopoly power*** The seller must be a monopolist or, at least, must possess some degree of monopoly power, that is, some ability to control output and price.
- ***Market segregation*** At relatively low cost to itself, the seller must be able to segregate buyers into distinct classes, each of which has a different willingness or ability to pay for the product. This separation of buyers is usually based on different price elasticities of demand, as the examples below will make clear.
- ***No resale*** The original purchaser cannot resell the product or service. If buyers in the low-price segment of the market could easily resell in the high-price segment, the monopolist's price-discrimination strategy would create competition in the high-price segment. This competition would reduce the price in the high-price segment and undermine the monopolist's price-discrimination policy. This condition suggests that service industries such as the transportation industry or legal and medical services, where resale is impossible, are good candidates for price discrimination.

Examples of Price Discrimination

Price discrimination is widely practiced in the U.S. economy. For example, we noted in Chapter 6's Last Word that airlines charge high fares to business travelers, whose demand for travel is inelastic, and offer lower, highly restricted, nonrefundable fares to attract vacationers and others whose demands are more elastic.

Electric utilities frequently segment their markets by end uses, such as lighting and heating. The absence of reasonable lighting substitutes means that the demand for electricity for illumination is inelastic and that the price per kilowatt-hour for such use is high. But the availability of natural gas and petroleum for heating makes the demand for electricity for this purpose less inelastic and the price lower.

Movie theaters and golf courses vary their charges on the basis of time (for example, higher evening and weekend rates) and age (for example, lower rates for children, senior discounts). Railroads vary the rate charged per ton-mile of freight according to the market value of the product being shipped. The shipper of 10 tons of television sets or refrigerators is charged more than the shipper of 10 tons of gravel or coal.

The issuance of discount coupons, redeemable at purchase, is a form of price discrimination. It enables firms to give price discounts to their most price-sensitive customers who have elastic demand. Less price-sensitive consumers who have less elastic demand are not as likely to take the time to clip and redeem coupons. The firm thus makes a larger profit than if it had used a single-price, no-coupon strategy.

Finally, price discrimination often occurs in international trade. A Russian aluminum producer, for example, might sell aluminum for less in the United States than in Russia. In the United States, this seller faces an elastic demand because several substitute suppliers are available. But in Russia, where the manufacturer dominates the market and trade barriers impede imports, consumers have fewer choices and thus demand is less elastic.

Graphical Analysis

Figure 12.8 demonstrates graphically the most frequently seen form of price discrimination—charging different prices to different classes of buyers. The two side-to-side graphs are for a single pure monopolist selling its product, say, software, in two segregated parts of the market. Figure 12.8a illustrates demand for software by small-business customers; Figure 12.8b, the demand for software by students. Student versions of the software are identical to the versions sold to businesses but are available (1 per person) only to customers with a student ID. Presumably, students have lower ability to pay for the software and are charged a discounted price.

The demand curve D_b in the graph to the left indicates a relatively inelastic demand for the product on the part of business customers. The demand curve D_s in the right-hand graph reflects the more elastic demand of students. The marginal revenue curves (MR_b and MR_s) lie below their respective demand curves, reflecting the demand–marginal revenue relationship previously described.

For visual clarity we have assumed that average total cost (ATC) is constant. Therefore marginal cost (MC) equals average total cost (ATC) at all quantities of output. These costs are the same for both versions of the software and therefore appear as the identical straight lines labeled "MC = ATC."

What price will the pure monopolist charge to each set of customers? Using the MR = MC rule for profit maximization, the firm will offer Q_b units of the software for sale to small businesses. It can sell that profit-maximizing output by charging price P_b. Again using the MR = MC rule, the monopolist will offer Q_s units of software to students. To sell those Q_s units, the firm will charge students the lower price P_s.

WORKED PROBLEMS

W12.2

Price discrimination

CONSIDER THIS . . .

Some Price Differences at the Ballpark

Take me out to the ball game . . .

Buy me some peanuts and Cracker Jack . . .

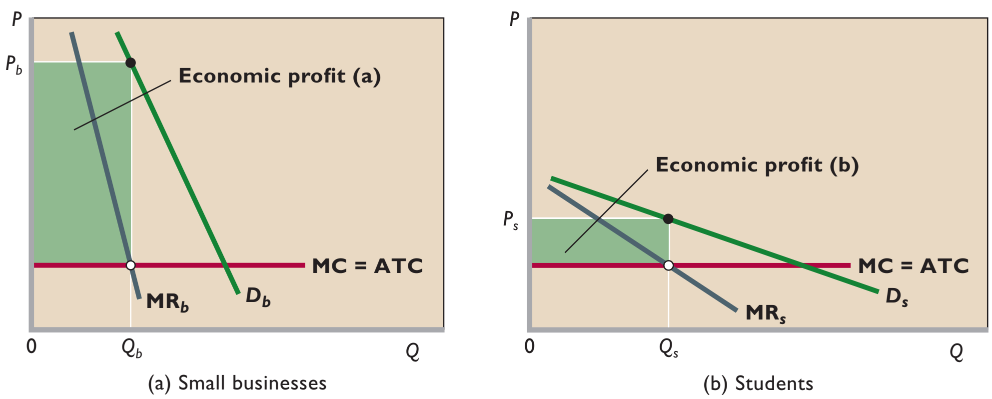
Professional baseball teams earn substantial revenues through ticket sales. To maximize profit, they offer significantly lower ticket prices for children (whose demand is elastic) than for adults (whose demand is inelastic). This discount may be as much as 50 percent.

If this type of price discrimination increases revenue and profit, why don't teams also price discriminate at the concession stands? Why don't they offer half-price hot dogs, soft drinks, peanuts, and Cracker Jack to children?

The answer involves the three requirements for successful price discrimination. All three requirements are met for game tickets: (1) The team has monopoly power; (2) it can segregate ticket buyers by age group, each group having a different elasticity of demand; and (3) children cannot resell their discounted tickets to adults.

It's a different situation at the concession stands. Specifically, the third condition is *not* met. If the team had dual prices, it could not prevent the exchange or "resale" of the concession goods from children to adults. Many adults would send children to buy food and soft drinks for them: "Here's some money, Billy. Go buy *six* hot dogs." In this case, price discrimination would reduce, not increase, team profit. Thus, children and adults are charged the same high prices at the concession stands. (These prices are high relative to those for the same goods at the local convenience store because the stadium sellers have a captive audience and thus considerable monopoly power.)

FIGURE 12.8 Price discrimination to different groups of buyers. The price-discriminating monopolist represented here maximizes its total profit by dividing the market into two segments based on differences in elasticity of demand. It then produces and sells the MR = MC output in each market segment. (For visual clarity, average total cost (ATC) is assumed to be constant. Therefore, MC equals ATC at all output levels.) (a) The price-discriminating monopolist charges a high price (here P_b) to small-business customers because they have a relatively inelastic demand curve for the product. (b) The firm charges a low price (here P_s) to students because their demand curve is relatively elastic. The firm's total profit from using price discrimination (here, the sum of the two green rectangles) exceeds the profit (not shown) that would have occurred if the monopolist had charged the same price to all customers.



Firms engage in price discrimination because it enhances their profit. The numbers (not shown) behind the curves in Figure 12.8 would clearly reveal that the sum of the two profit rectangles shown in green exceeds the single profit rectangle the firm would obtain from a single monopoly price. How do consumers fare? In this case, students clearly benefit by paying a lower price than they would if the firm charged a single monopoly price; in contrast, the price discrimination results in a higher price for business customers. Therefore, compared to the single-price situation, students buy more of the software and small businesses buy less.

Such price discrimination is widespread in the economy and is illegal only when it is part of a firm's strategy to lessen or eliminate competition. We will discuss illegal price discrimination in Chapter 19, which covers antitrust policy.

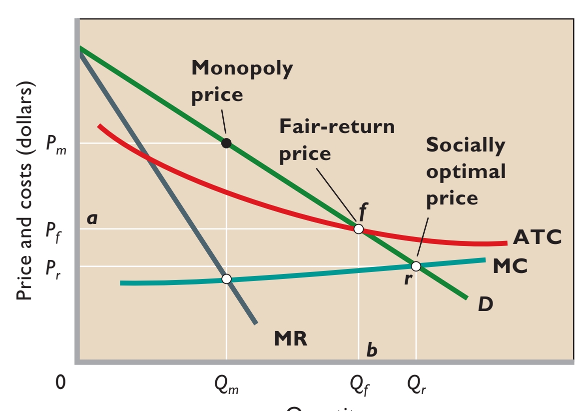
Regulated Monopoly

LO12.7 Distinguish between the monopoly price, the socially optimal price, and the fair-return price of a government-regulated monopoly.

Natural monopolies traditionally have been subject to *rate regulation* (price regulation), although the recent trend has been to deregulate wherever competition seems possible. For example, long-distance telephone calls, natural gas distribution, wireless communications, cable television, and long-distance electricity transmission have been, to one degree or another, deregulated over the past several decades. And regulators in some states are beginning to allow new entrants to compete with existing local telephone and electricity providers. Nevertheless, state and local regulatory commissions still regulate the prices that most local natural gas distributors, regional telephone companies, and local electricity suppliers can charge. These locally regulated monopolies are commonly called "public utilities."

Let's consider the regulation of a local natural monopoly. Our example will be a single firm that is the only seller of natural gas in the town of Springfield. Figure 12.9

FIGURE 12.9 Regulated monopoly. The socially optimal price P_r, found where D and MC intersect, will result in an efficient allocation of resources but may entail losses to the monopoly. The fair-return price P_f will allow the monopolist to break even but will not fully correct the underallocation of resources.



shows the demand and the long-run cost curves facing our firm. Because of extensive economies of scale, the demand curve cuts the natural monopolist's long-run average-total-cost curve at a point where that curve is still falling. It would be inefficient to have several firms in this industry because each would produce a much smaller output, operating well to the left on the long-run average-total-cost curve. In short, each firm's lowest average total cost would be substantially higher than that of a single firm. So efficient, lowest-cost production requires a single seller.

We know by application of the MR = MC rule that Q_m and P_m are the profit-maximizing output and price that an unregulated monopolist would choose. Because price exceeds average total cost at output Q_m, the monopolist enjoys a substantial economic profit. Furthermore, price exceeds marginal cost, indicating an underallocation of resources to this product or service. Can government regulation bring about better results from society's point of view?

Socially Optimal Price: P = MC

One sensible goal for regulators would be to get the monopoly to produce the allocatively efficient output level. For our monopolist in Figure 12.9, this is output level Q_r, determined by where the demand curve D intersects the MC curve. Q_r is the allocatively efficient output level because for each unit of output up to Q_r, the demand curve lies above the MC curve, indicating that for all of these units marginal benefits exceed marginal costs.

But how can the regulatory commission actually motivate the monopoly to produce this output level? The trick is to set the regulated price P_r at a level such that the monopoly will be led by its profit-maximizing rule to voluntarily produce the allocatively efficient level of output. To see how this works, note that because the monopoly will receive the regulated price P_r for all units that it sells, P_r becomes the monopoly's marginal revenue per unit. Thus, the monopoly's MR curve becomes the horizontal white line moving rightward from price P_r on the vertical axis.

The monopoly will at this point follow its usual rule for maximizing profits or minimizing losses: It will produce where marginal revenue equals marginal cost. As a result, the monopoly will produce where the horizontal white MR (= P_r) line intersects the MC curve at point r. That is, the monopoly will end up producing the socially optimal output Q_r not because it is socially minded but because Q_r happens to be the output that either maximizes profits or minimizes losses when the firm is forced by the regulators to sell all units at the regulated price P_r.

The regulated price P_r that achieves allocative efficiency is called the **socially optimal price.** Because it is determined by where the MC curve intersects the demand curve, this type of regulation is often summarized by the equation P = MC.

Fair-Return Price: P = ATC

The socially optimal price suffers from a potentially fatal problem. P_r may be so low that average total costs are not covered, as is the case in Figure 12.9. In such situations, forcing the socially optimal price on the regulated monopoly would result in short-run losses and long-run exit. In our example, Springfield would be left without a gas company and its citizens without gas.

What can be done to rectify this problem? One option is to provide a public subsidy to cover the loss that the socially optimal price would entail. Another possibility is to condone price discrimination, allow the monopoly to charge some customers prices above P_r, and hope that the additional revenue that the monopoly gains from price discrimination will be enough to permit it to break even.

In practice, regulatory commissions in the United States have often pursued a third option that abandons the goal of producing every unit for which marginal benefits exceed marginal costs but that guarantees that regulated monopolies will be able to break even and continue in operation. Under this third option, regulators set a regulated price that is high enough for monopolists to break even and continue in operation. This price has come to be referred to as a **fair-return price** because of a ruling in which the Supreme Court held that regulatory agencies must permit regulated utility owners to enjoy a "fair return" on their investments.

In practice, a fair return is equal to a normal profit. That is, a fair return is an accounting profit equal in size to what the owners of the monopoly would on average receive if they entered another type of business.

The regulator determines the fair-return price P_f by where the average total cost curve intersects the demand curve at point f. As we will explain, setting the regulated price at this level will cause the monopoly to produce Q_f units while guaranteeing that it will break even and not wish to exit the industry. To see why the monopoly will voluntarily produce Q_f units, note that because the monopoly will receive P_f dollars for each unit it sells, its marginal revenue per unit becomes P_f dollars so that the horizontal line moving rightward from P_f on the vertical axis becomes the regulated monopoly's MR curve. Because this horizontal MR curve is always higher than the monopoly's MC

LAST WORD

Monopoly Power in the Internet Age

Network Effects and Economies of Scale Have Driven the Monopolistic Growth of Several Internet Giants, Including Google, Facebook, and Amazon.

In the early 1990s, when the Internet was young, many analysts predicted that it would foster pure competition across a wide range of activities. Because the Internet allowed any user to publish text and images that could be read for free by any other user, they assumed that the Internet would create a level playing field for all types of media, communications, and commerce.

These predictions turned out to be wrong. One mistake was in not understanding that in a world awash in information, finding what you want becomes a huge problem. When the Internet started, there was no directory and there were no search engines. So it was nearly impossible to find what you were looking for.

Google solved that problem by creating the first effective search engine. Thanks to Google, people could easily locate what they were looking for. But this meant that anyone wishing to be found was now dependent on Google or some other search engine to be found.

If you were an advertiser, you would want to spend your money placing keyword ads on the most popular search engine so that your ads would reach as many potential customers as possible. And if you were a customer who found ads helpful in finding what you were looking for, you would also want to utilize the most popular search engine so that you could be exposed to the greatest number of helpful ads. Thus, Google

curve, it is obvious that marginal revenues will exceed marginal costs for every possible level of output shown in Figure 12.9. Thus, the monopoly should be willing to supply whatever quantity of output is demanded by consumers at the regulated price P_f. That quantity is, of course, given by the demand curve. At price P_f consumers will demand exactly Q_f units. Thus, by setting the regulated price at P_f, the regulator gets the monopoly to voluntarily supply exactly Q_f units.

Even better, the regulator also guarantees that the monopoly firm will earn exactly a normal profit. This can be seen in Figure 12.9 by noting that the rectangle 0*afb* is equal to both the monopoly's total cost and its total revenue. Its economic profit is therefore equal to zero, implying that it must be earning a normal accounting profit for its owners.

One final point about allocative efficiency: By choosing the fair-return price P_f, the regulator leads the monopoly to produce Q_f units. This is less than the socially optimal quantity Q_r, but still more than the Q_m units that the monopolist would produce if left unregulated. So while fair-return pricing does not lead to full allocative efficiency, it is still an improvement on what the monopoly would do if left to its own devices.

Dilemma of Regulation

Comparing results of the socially optimal price (P = MC) and the fair-return price (P = ATC) suggests a policy dilemma, sometimes termed the *dilemma of regulation*. When its price is set to achieve the most efficient allocation of resources (P = MC), the regulated monopoly is likely to suffer losses. Survival of the firm would presumably depend on permanent public subsidies out of tax revenues. On the other hand, although a fair-return price (P = ATC) allows the monopolist to cover costs, it only partially resolves the underallocation of resources that the unregulated monopoly price would foster. Despite this dilemma, regulation can improve on the results of monopoly from the social point of view. Price regulation (even at the quickly came to dominate search as the result of network effects by which the value of Google to any one particular user increased with the total number of users.

Those network effects also created a barrier to entry that protects Google from competitors because both those searching for information and as well as those wanting to provide it have an interest in sticking with whatever search engine has the most users. There are in fact many smaller search engines, but nobody wants to use them very much for the simple reason that almost nobody else is using them. Consequently, Google controls nearly 70 percent of the U.S. search market and receives about 75 percent of the revenue generated by search ads.

The network effects that help Google dominate search also drive the dominance that just a handful of firms hold over other parts of the Internet. Consider Facebook. It is a well-run Web site with lots of interesting things to do. But most people come back for the wall posts and other content generated by fellow users. If there were no fellow users, there would be little content and little reason to visit the site.

That makes it hard for smaller social-networking sites to compete with Facebook. If interacting with others is the whole point of joining a networking site, why would you want to join a site that has very few people to interact with? As a result, Facebook has come to dominate social media. With over a billion users, it enjoys the largest network effect and grows even bigger thanks to already being big.

The early predictions that the Internet would create a level playing field for all types of media, communications, and commerce have also been doomed by economies of scale. Consider Amazon. To the public, Amazon is the world's largest online retailer, with over $50 billion in annual sales. But behind the scenes, its success is driven by two activities that each enjoys massive economies of scale: data and logistics.

In terms of data, Amazon runs some of the world's largest server farms. These giant buildings are stacked top to bottom with tens of thousands of networked computers that store customer data, process payments, and keep track of inventory. The cost of building and running these server farms runs into the billions of dollars each year—including massive electricity bills. But because a larger server farm generates a lower cost per sale than a smaller server farm, Amazon enjoys economies of scale that allow it to undersell any rival operating on a smaller scale with smaller server farms.

The story with logistics is much the same. Amazon operates dozens of massive distribution warehouses that benefit from economies of scale because a warehouse that is twice as big costs less than twice as much to operate.

We should note, however, that Google, Facebook, and Amazon are not full-on monopolies. Each faces robust competition. While network effects and economies of scale benefit them greatly, those factors are not strong enough to guarantee them permanent dominance or even large profits. Amazon's 2011 profit was only 1.3 percent.

fair-return price) can simultaneously reduce price, increase output, and reduce the economic profit of monopolies.

That said, we need to provide an important caution: "Fair-price" regulation of monopoly looks rather simple in theory but is amazingly complex in practice. In the actual economy, rate regulation is accompanied by large, expensive rate-setting bureaucracies and maze-like sets of procedures. Also, rate decisions require extensive public input via letters and through public hearings. Rate decisions are subject to lengthy legal challenges. Further, because regulatory commissions must set prices sufficiently above costs to create fair returns, regulated monopolists have little incentive to minimize average total costs. When these costs creep up, the regulatory commissions must set higher prices.

Regulated firms therefore are noted for higher-than-competitive wages, more managers and staff than necessary, nicer-than-typical office buildings, and other forms of X-inefficiency. These inefficiencies help explain the trend of federal, state, and local governments abandoning price regulation where the possibility of competition looks promising.

QUICK REVIEW 12.3

- Price discrimination occurs when a firm sells a product at different prices that are not based on cost differences.
- The conditions necessary for price discrimination are (a) monopoly power, (b) the ability to segregate buyers on the basis of demand elasticities, and (c) the inability of buyers to resell the product.
- Compared with single pricing by a monopolist, perfect price discrimination results in greater profit and greater output. Many consumers pay higher prices, but other buyers pay prices below the single price.
- Monopoly price can be reduced and output increased through government regulation.
- The socially optimal price (P = MC) achieves allocative efficiency but may result in losses; the fair-return price (P = ATC) yields a normal profit but fails to achieve allocative efficiency.

SUMMARY

LO12.1 List the characteristics of pure monopoly.

A pure monopolist is the sole producer of a commodity for which there are no close substitutes.

LO12.2 List and explain the barriers to entry that shield pure monopolies from competition.

The existence of pure monopoly and other imperfectly competitive market structures is explained by barriers to entry in the form of (*a*) economies of scale, (*b*) patent ownership and research, (*c*) ownership or control of essential resources, and (*d*) pricing and other strategic behavior.

LO12.3 Explain how demand is seen by a pure monopoly.

The pure monopolist's market situation differs from that of a competitive firm in that the monopolist's demand curve is downsloping, causing the marginal-revenue curve to lie below the demand curve. Like the competitive seller, the pure monopolist will maximize profit by equating marginal revenue and marginal cost. Barriers to entry may permit a monopolist to acquire economic profit even in the long run. However, (*a*) the monopolist does not charge "the highest price possible"; (*b*) the price that yields maximum total profit to the monopolist rarely coincides with the price that yields maximum unit profit; (*c*) high costs and a weak demand may prevent the monopolist from realizing any profit at all; and (*d*) the monopolist avoids the inelastic region of its demand curve.

LO12.4 Explain how a pure monopoly sets its profit-maximizing output and price.

With the same costs, the pure monopolist will find it profitable to restrict output and charge a higher price than would sellers in a purely competitive industry. This restriction of output causes resources to be misallocated, as is evidenced by the fact that price exceeds marginal cost in monopolized markets. Monopoly creates an efficiency loss (or deadweight loss) for society.

Monopoly transfers income from consumers to monopolists because a monopolist can charge a higher price than would a purely competitive firm with the same costs. So monopolists in effect levy a "private tax" on consumers and, if demand is strong enough, obtain substantial economic profits.

LO12.5 Discuss the economic effects of monopoly.

The costs monopolists and competitive producers face may not be the same. On the one hand, economies of scale may make lower unit costs available to monopolists but not to competitors. Also, pure monopoly may be more likely than pure competition to reduce costs via technological advance because of the monopolist's ability to realize economic profit, which can be used to finance research. On the other hand, X-inefficiency—the failure to produce with the least costly combination of inputs—is more common among monopolists than among competitive firms. Also, monopolists may make costly expenditures to maintain monopoly privileges that are conferred by government. Finally, the blocked entry of rival firms weakens the monopolist's incentive to be technologically progressive.

LO12.6 Describe why a monopolist might prefer to charge different prices in different markets.

A monopolist can increase its profit by practicing price discrimination, provided (*a*) it can segregate buyers on the basis of elasticities of demand and (*b*) its product or service cannot be readily transferred between the segregated markets.

LO12.7 Distinguish between the monopoly price, the socially optimal price, and the fair-return price of a government-regulated monopoly.

Price regulation can be invoked to eliminate wholly or partially the tendency of monopolists to underallocate resources and to earn economic profits. The socially optimal price is determined where the demand and marginal-cost curves intersect; the fair-return price is determined where the demand and average-total-cost curves intersect.

TERMS AND CONCEPTS

pure monopoly
barriers to entry
simultaneous consumption
network effects
X-inefficiency
rent-seeking behavior
price discrimination
socially optimal price
fair-return price

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. "No firm is completely sheltered from rivals; all firms compete for consumer dollars. If that is so, then pure monopoly does not exist." Do you agree? Explain. How might you use Chapter 6's concept of cross elasticity of demand to judge whether monopoly exists? **LO12.1**

2. Discuss the major barriers to entry into an industry. Explain how each barrier can foster either monopoly or oligopoly. Which barriers, if any, do you feel give rise to monopoly that is socially justifiable? **LO12.2**
3. How does the demand curve faced by a purely monopolistic seller differ from that confronting a purely competitive firm? Why does it differ? Of what significance is the difference? Why is the pure monopolist's demand curve not perfectly inelastic? **LO12.3**
4. Assume that a pure monopolist and a purely competitive firm have the same unit costs. Contrast the two with respect to (*a*) price, (*b*) output, (*c*) profits, (*d*) allocation of resources, and (*e*) impact on income transfers. Since both monopolists and competitive firms follow the MC = MR rule in maximizing profits, how do you account for the different results? Why might the costs of a purely competitive firm and those of a monopolist be different? What are the implications of such a cost difference? **LO12.5**
5. Critically evaluate and explain each statement: **LO12.5**
   a. Because they can control product price, monopolists are always assured of profitable production by simply charging the highest price consumers will pay.
   b. The pure monopolist seeks the output that will yield the greatest per-unit profit.
   c. An excess of price over marginal cost is the market's way of signaling the need for more production of a good.
   d. The more profitable a firm, the greater its monopoly power.
   e. The monopolist has a pricing policy; the competitive producer does not.
   f. With respect to resource allocation, the interests of the seller and of society coincide in a purely competitive market but conflict in a monopolized market.
6. Assume a monopolistic publisher has agreed to pay an author 10 percent of the total revenue from the sales of a text. Will the author and the publisher want to charge the same price for the text? Explain. **LO12.5**
7. U.S. pharmaceutical companies charge different prices for prescription drugs to buyers in different nations, depending on elasticity of demand and government-imposed price ceilings. Explain why these companies, for profit reasons, oppose laws allowing reimportation of drugs to the United States. **LO12.6**
8. Explain verbally and graphically how price (rate) regulation may improve the performance of monopolies. In your answer distinguish between (*a*) socially optimal (marginal-cost) pricing and (*b*) fair-return (average-total-cost) pricing. What is the "dilemma of regulation"? **LO12.7**
9. It has been proposed that natural monopolists should be allowed to determine their profit-maximizing outputs and prices and then government should tax their profits away and distribute them to consumers in proportion to their purchases from the monopoly. Is this proposal as socially desirable as requiring monopolists to equate price with marginal cost or average total cost? **LO12.7**
10. **LAST WORD** How do network effects help Facebook fend off smaller social-networking rivals? Could an online retailer doing half as much business compete on an equal footing with Amazon in terms of costs? Explain.

REVIEW QUESTIONS

1. Which of the following could explain why a firm is a monopoly? Select one or more answers from the choices shown. **LO12.2**
   a. Patents.
   b. Economies of scale.
   c. Inelastic demand.
   d. Government licenses.
   e. Downsloping market demand.
2. The MR curve of a perfectly competitive firm is horizontal. The MR curve of a monopoly firm is: **LO12.3**
   a. Horizontal, too.
   b. Upsloping.
   c. Downsloping.
   d. It depends.
3. Use the nearby demand schedule to calculate total revenue and marginal revenue at each quantity. Plot the demand, total-revenue, and marginal-revenue curves, and explain the relationships between them. Explain why the marginal revenue of the fourth unit of output is $3.50, even though its price is $5. Use Chapter 6's total-revenue test for price elasticity to designate the elastic and inelastic segments of your graphed demand curve. What generalization can you make as to the relationship between marginal revenue and elasticity of demand? Suppose the marginal cost of successive units of output was zero. What output would the profit-seeking firm produce? Finally, use your analysis to explain why a monopolist would never produce in the inelastic region of demand. **LO12.3**

Price (*P*)	Quantity Demanded (*Q*)	Price (*P*)	Quantity Demanded (*Q*)
$7.00	0	$4.50	5
6.50	1	4.00	6
6.00	2	3.50	7
5.50	3	3.00	8
5.00	4	2.50	9

4. How often do *perfectly competitive* firms engage in price discrimination? **LO12.6**
   a. Never.
   b. Rarely.
   c. Often.
   d. Always.

5. Suppose that a monopolist can segregate his buyers into two different groups to which he can charge two different prices. In order to maximize profit, the monopolist should charge a higher price to the group that has: **LO12.6**
   a. The higher elasticity of demand.
   b. The lower elasticity of demand.
   c. Richer members.
6. The socially optimal price ($P = MC$) is socially optimal because: **LO12.7**
   a. It reduces the monopolist's profit.
   b. It yields a normal profit.
   c. It minimizes ATC.
   d. It achieves allocative efficiency.
7. The main problem with imposing the socially optimal price ($P = MC$) on a monopoly is that the socially optimal price: **LO12.7**
   a. May be so low that the regulated monopoly can't break even.
   b. May cause the regulated monopoly to engage in price discrimination.
   c. May be higher than the monopoly price.

PROBLEMS

1. Suppose a pure monopolist is faced with the demand schedule shown below and the same cost data as the competitive producer discussed in problem 4 at the end of Chapter 10. Calculate the missing total-revenue and marginal-revenue amounts, and determine the profit-maximizing price and profit-maximizing output for this monopolist. What is the monopolist's profit? Verify your answer graphically and by comparing total revenue and total cost. **LO12.4**

Price	Quantity Demanded	Total Revenue	Marginal Revenue
$115	0	$______	
100	1	______	$______
83	2	______	______
71	3	______	______
63	4	______	______
55	5	______	______
48	6	______	______
42	7	______	______
37	8	______	______
33	9	______	______
29	10	______	______

2. Suppose that a price-discriminating monopolist has segregated its market into two groups of buyers. The first group is described by the demand and revenue data that you developed for problem 1. The demand and revenue data for the second group of buyers is shown in the following table. Assume that MC is $13 in both markets and MC = ATC at all output levels. What price will the firm charge in each market? Based solely on these two prices, which market has the higher price elasticity of demand? What will be this monopolist's total economic profit? **LO12.6**

Price	Quantity Demanded	Total Revenue	Marginal Revenue
$71	0	$ 0	
63	1	63	$63
55	2	110	47
48	3	144	34
42	4	168	24
37	5	185	17
33	6	198	13
29	7	203	5

3. Assume that the most efficient production technology available for making vitamin pills has the cost structure given in the following table. Note that output is measured as the number of bottles of vitamins produced per day and that costs include a normal profit. **LO12.6**

Output	TC	MC
25,000	$100,000	$0.50
50,000	150,000	1.00
75,000	187,500	2.50
100,000	275,500	3.00

   a. What is ATC per unit for each level of output listed in the table?
   b. Is this a decreasing-cost industry? (Answer yes or no).
   c. Suppose that the market price for a bottle of vitamins is $2.50 and that at that price the total market quantity demanded is 75,000,000 bottles. How many firms will there be in this industry?
   d. Suppose that, instead, the market quantity demanded at a price of $2.50 is only 75,000. How many firms do you expect there to be in this industry?
   e. Review your answers to parts *b*, *c*, and *d*. Does the level of demand determine this industry's market structure?

4. A new production technology for making vitamins is invented by a college professor who decides not to patent it. Thus, it is available for anybody to copy and put into use. The TC per bottle for production up to 100,000 bottles per day is given in the following table. **LO12.6**

Output	TC
25,000	$50,000
50,000	70,000
75,000	75,000
100,000	80,000

   a. What is ATC for each level of output listed in the table?
   b. Suppose that for each 25,000-bottle-per-day increase in production above 100,000 bottles per day, TC increases by $5,000 (so that, for instance, 125,000 bottles per day would generate total costs of $85,000 and 150,000 bottles per day would generate total costs of $90,000). Is this a decreasing-cost industry?
   c. Suppose that the price of a bottle of vitamins is $1.33 and that at that price the total quantity demanded by consumers is 75,000,000 bottles. How many firms will there be in this industry?
   d. Suppose that, instead, the market quantity demanded at a price of $1.33 is only 75,000. How many firms do you expect there to be in this industry?
   e. Review your answers to parts *b*, *c*, and *d*. Does the level of demand determine this industry's market structure?
   f. Compare your answer to part *d* of this problem with your answer to part d of problem 3. Do both production technologies show constant returns to scale?

5. Suppose you have been tasked with regulating a single monopoly firm that sells 50-pound bags of concrete. The firm has fixed costs of $10 million per year and a variable cost of $1 per bag no matter how many bags are produced. **LO12.7**
   a. If this firm kept on increasing its output level, would ATC per bag ever increase? Is this a decreasing-cost industry?
   b. If you wished to regulate this monopoly by charging the socially optimal price, what price would you charge? At that price, what would be the size of the firm's profit or loss? Would the firm want to exit the industry?
   c. You find out that if you set the price at $2 per bag, consumers will demand 10 million bags. How big will the firm's profit or loss be at that price?
   d. If consumers instead demanded 20 million bags at a price of $2 per bag, how big would the firm's profit or loss be?
   e. Suppose that demand is perfectly inelastic at 20 million bags, so that consumers demand 20 million bags no matter what the price is. What price should you charge if you want the firm to earn only a fair rate of return? Assume as always that TC includes a normal profit.

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **13**

Monopolistic Competition and Oligopoly

Learning Objectives

LO13.1 List the characteristics of monopolistic competition.

LO13.2 Explain why monopolistic competitors earn only a normal profit in the long run.

LO13.3 Explain why monopolistic competition delivers neither productive nor allocative efficiency.

LO13.4 Relate how the ability of monopolistic competition to deliver product differentiation helps to compensate for its failure to deliver economic efficiency.

LO13.5 Describe the characteristics of oligopoly.

LO13.6 Discuss how game theory relates to oligopoly.

LO13.7 Explain the three main models of oligopoly pricing and output: kinked-demand theory, collusive pricing, and price leadership.

LO13.8 Contrast the potential positive and negative effects of advertising.

LO13.9 Discuss the efficiency of oligopoly from society's standpoint and whether it is more or less efficient than monopoly.

LO13.10 (Appendix) Utilize additional game-theory terminology and applications.

In the United States, most industries have a market structure that falls somewhere between the two poles of pure competition and pure monopoly. To begin with, most real-world industries have fewer than the large number of producers required for pure competition but more than the single producer that defines pure monopoly. In addition, most firms in most industries have both distinguishable rather than standardized products as well as some discretion over the prices they charge. As a result, competition often occurs on the basis of price, quality, location, service, and advertising. Finally, entry to most real-world industries ranges from easy to very difficult but is rarely completely blocked.

This chapter examines two models that more closely approximate these widespread industry structures. You will discover that *monopolistic competition* mixes a small amount of monopoly power with a large amount of competition. *Oligopoly,* in contrast, blends a large amount of monopoly power with both considerable rivalry among existing firms and the threat of increased future competition due to foreign firms and new technologies. (You should quickly review Table 10.1, page 221, at this point.)

Monopolistic Competition

LO13.1 List the characteristics of monopolistic competition.

Let's begin by examining **monopolistic competition,** which is characterized by (1) a relatively large number of sellers, (2) differentiated products (often promoted by heavy advertising), and (3) easy entry to, and exit from, the industry. The first and third characteristics provide the "competitive" aspect of monopolistic competition; the second characteristic provides the "monopolistic" aspect. In general, however, monopolistically competitive industries are much more competitive than they are monopolistic.

ORIGIN OF THE IDEA

O13.1
Monopolistic competition

Relatively Large Number of Sellers

Monopolistic competition is characterized by a fairly large number of firms, say, 25, 35, 60, or 70, not by the hundreds or thousands of firms in pure competition. Consequently, monopolistic competition involves:

- ***Small market shares*** Each firm has a comparatively small percentage of the total market and consequently has limited control over market price.
- ***No collusion*** The presence of a relatively large number of firms ensures that collusion by a group of firms to restrict output and set prices is unlikely.
- ***Independent action*** With numerous firms in an industry, there is no feeling of interdependence among them; each firm can determine its own pricing policy without considering the possible reactions of rival firms. A single firm may realize a modest increase in sales by cutting its price, but the effect of that action on competitors' sales will be nearly imperceptible and will probably trigger no response.

Differentiated Products

In contrast to pure competition, in which there is a standardized product, monopolistic competition is distinguished by **product differentiation.** Monopolistically competitive firms turn out variations of a particular product. They produce products with slightly different physical characteristics, offer varying degrees of customer service, provide varying amounts of locational convenience, or proclaim special qualities, real or imagined, for their products.

Let's examine these aspects of product differentiation in more detail.

Product Attributes Product differentiation may entail physical or qualitative differences in the products themselves. Real differences in functional features, materials, design, and workmanship are vital aspects of product differentiation. Personal computers, for example, differ in terms of storage capacity, speed, graphic displays, and included software. There are dozens of competing principles of economics textbooks that differ in content, organization, presentation and readability, pedagogical aids, and graphics and design. Most cities have a variety of retail stores selling men's and women's clothes that differ greatly in styling, materials, and quality of work. Similarly, one pizza place may feature thin-crust Neapolitan style pizza, while another may tout its thick-crust Chicago-style pizza.

Service Service and the conditions surrounding the sale of a product are forms of product differentiation too. One shoe store may stress the fashion knowledge and helpfulness of its clerks. A competitor may leave trying on shoes and carrying them to the register to its customers but feature lower prices. Customers may prefer one-day over three-day dry cleaning of equal quality. The prestige appeal of a store, the courteousness and helpfulness of clerks, the firm's reputation for servicing or exchanging its products, and the credit it makes available are all service aspects of product differentiation.

Location Products may also be differentiated through the location and accessibility of the stores that sell them. Small convenience stores manage to compete with large supermarkets, even though these minimarts have a more limited range of products and charge higher prices. They compete mainly on the basis of location—being close to customers and situated on busy streets. A motel's proximity to an interstate highway gives it a locational advantage that may enable it to charge a higher room rate than nearby motels in less convenient locations.

Brand Names and Packaging Product differentiation may also be created through the use of brand names and trademarks, packaging, and celebrity connections. Most aspirin tablets are very much alike, but many headache sufferers believe that one brand—for example, Bayer, Anacin, or Bufferin—is superior and worth a higher price than a generic substitute. A celebrity's name associated with watches, perfume, or athletic shoes may enhance the appeal of those products for some buyers. Many customers prefer one style of ballpoint pen to another. Packaging that touts "natural spring" bottled water may attract additional customers.

Some Control over Price Despite the relatively large number of firms, monopolistic competitors do have some control over their product prices because of product differentiation. If consumers prefer the products of specific sellers, then within limits they will pay more to satisfy their preferences. Sellers and buyers are not linked randomly, as in a purely competitive market. But the monopolistic competitor's control over price is quite limited since there are numerous potential substitutes for its product.

Easy Entry and Exit

Entry into monopolistically competitive industries is relatively easy compared to oligopoly or pure monopoly. Because monopolistic competitors are typically small firms, both absolutely and relatively, economies of scale are few and capital requirements are low. On the other hand, compared with pure competition, financial barriers may result from the need to develop and advertise a product that differs from rivals' products. Some firms have trade secrets relating to their products or hold trademarks on their brand names, making it difficult and costly for other firms to imitate them.

Exit from monopolistically competitive industries is relatively easy. Nothing prevents an unprofitable monopolistic competitor from holding a going-out-of-business sale and shutting down.

Advertising

The expense and effort involved in product differentiation would be wasted if consumers were not made aware of product differences. Thus, monopolistic competitors advertise their products, often heavily. The goal of product differentiation and advertising—so-called **nonprice competition**—is to make price less of a factor in consumer purchases and make product differences a greater factor. If successful, the firm's demand curve will shift to the right and will become less elastic.

Monopolistically Competitive Industries

Table 13.1 lists several manufacturing industries that approximate monopolistic competition. Economists measure the degree of industry concentration—the extent to which the largest firms account for the bulk of the industry's output—to identify monopolistically competitive (versus oligopolistic) industries. Two such measures are the four-firm concentration ratio and the Herfindahl index. They are listed in columns 2 and 3 of the table.

A **four-firm concentration ratio,** expressed as a percentage, is the ratio of the output (sales) of the four largest firms in an industry relative to total industry sales.

$$\text{Four-firm concentration ratio} = \frac{\text{Output of four largest firms}}{\text{Total output in the industry}}$$

Four-firm concentration ratios are very low in purely competitive industries in which there are hundreds or even thousands of firms, each with a tiny market share. In contrast, four-firm ratios are high in oligopoly and pure monopoly. Industries in which the largest four firms account for 40 percent or more of the market are generally considered to be oligopolies. If the largest four firms account for less than 40 percent, they are likely to be monopolistically

TABLE 13.1 Percentage of Output Produced by Firms in Selected Low-Concentration U.S. Manufacturing Industries

(1) Industry	(2) Percentage of Industry Output* Produced by the Four Largest Firms	(3) Herfindahl Index for the Top 50 Firms	(1) Industry	(2) Percentage of Industry Output* Produced by the Four Largest Firms	(3) Herfindahl Index for the Top 50 Firms
Textile machinery	30	360	Wood trusses	15	102
Women's dresses	28	328	Metal stamping	14	88
Textile bags	28	318	Metal windows and doors	13	109
Plastic bags	27	299	Wood pallets	11	51
Ready-mix concrete	23	313	Sheet metal work	7	30
Jewelry	23	230	Signs	7	28
Asphalt paving	22	188	Stone products	7	23
Plastic pipe	21	187	Quick printing	4	8
Sawmills	15	98	Retail bakeries	4	7
Curtains and draperies	14	85	Bolts, nuts, and rivets	4	6

*As measured by value of shipments. Data are for 2007. See **www.census.gov/epcd/www/concentration.html**.

Source: Bureau of Census, *Census of Manufacturers,* 2007.

competitive. Observe that the four-firm concentration ratios in Table 13.1 range from 4 percent to 30 percent.

Published concentration ratios such as those in Table 13.1 are helpful in categorizing industries but must be used cautiously because the market shares (percentage of total sales) that they list are national in scope, whereas competition in many industries is often local in scope. As a result, some industries with low national concentration ratios are in fact substantially concentrated if one focuses on local markets.

As an example, the national four-firm concentration ratio for ready-mix concrete shown in Table 13.1 is only 23 percent. This suggests that ready-mix concrete is a monopolistically competitive industry. But the sheer bulk of ready-mix concrete and the fact that it "sets up" as it dries limits the relevant market to a specific town, city, or metropolitan area. In most of these local markets, only a few firms compete, not the numerous firms needed for monopolistic competition.

Column 3 of Table 13.1 lists a second measure of concentration: the **Herfindahl index.** This index is the sum of the squared percentage market shares of all firms in the industry. In equation form:

$$\text{Herfindahl index} = (\%S_1)^2 + (\%S_2)^2 + (\%S_3)^2 + \cdots + (\%S_n)^2$$

where $\%S_1$ is the percentage market share of firm 1, $\%S_2$ is the percentage market share of firm 2, and so on for each of the n total firms in the industry. By squaring the percentage market shares of all firms in the industry, the Herfindahl index purposely gives much greater weight to larger, and thus more powerful, firms than to smaller ones. For a purely competitive industry, the index would approach zero since each firm's market share—$\%S$ in the equation—is extremely small. In the case of a single-firm industry, the index would be at its maximum of 10,000 ($= 100^2$), indicating an industry with complete monopoly power.

We will discover later in this chapter that the Herfindahl index is important for assessing oligopolistic industries. But for now, the relevant generalization is that the lower the Herfindahl index, the greater is the likelihood that an industry is monopolistically competitive rather than oligopolistic. Column 3 of Table 13.1 lists the Herfindahl index (computed for the top 50 firms, not all the industry firms) for several industries. Note that the index values are decidedly closer to the bottom limit of the Herfindahl index—0—than to its top limit—10,000.

The numbers in Table 13.1 are for manufacturing industries. In addition, many retail establishments in metropolitan areas are monopolistically competitive, including grocery stores, gasoline stations, hair salons, dry cleaners, clothing stores, and restaurants. Also, many providers of professional services such as medical care, legal assistance, real estate sales, and basic bookkeeping are monopolistic competitors.

Price and Output in Monopolistic Competition

LO13.2 Explain why monopolistic competitors earn only a normal profit in the long run.

How does a monopolistic competitor decide on its price and output? To explain, we initially assume that each firm

KEY GRAPH

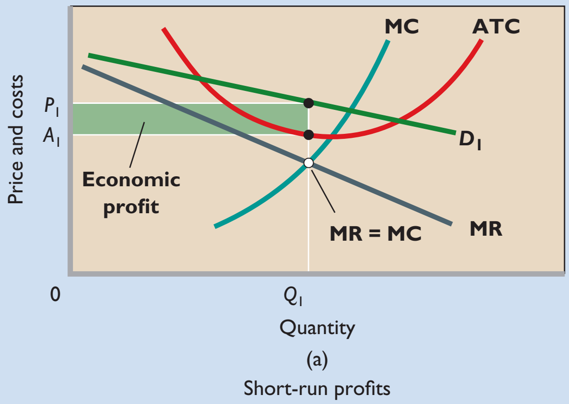
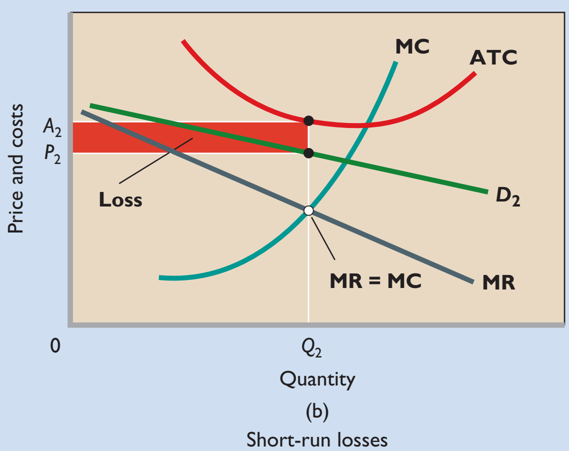
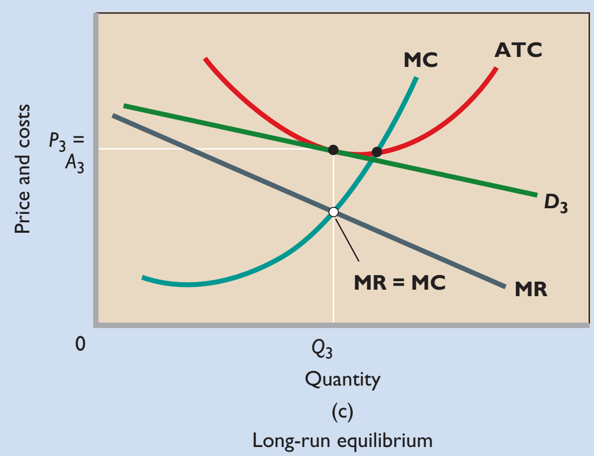


FIGURE 13.1 A monopolistically competitive firm: short run and long run. The monopolistic competitor maximizes profit or minimizes loss by producing the output at which MR = MC. The economic profit shown in (a) will induce new firms to enter, eventually eliminating economic profit. The loss shown in (b) will cause an exit of firms until normal profit is restored. After such entry and exit, the price will settle in (c) to where it just equals average total cost at the MR = MC output. At this price P_3 and output Q_3, the monopolistic competitor earns only a normal profit, and the industry is in long-run equilibrium.





QUICK QUIZ FOR FIGURE 13.1

1. Price exceeds MC in:
   - **a.** graph (a) only.
   - **b.** graph (b) only.
   - **c.** graphs (a) and (b) only.
   - **d.** graphs (a), (b), and (c).
2. Price exceeds ATC in:
   - **a.** graph (a) only.
   - **b.** graph (b) only.
   - **c.** graphs (a) and (b) only.
   - **d.** graphs (a), (b), and (c).
3. The firm represented by Figure 13.1c is:
   - **a.** making a normal profit.
   - **b.** incurring a loss.
   - **c.** producing at the same level of output as a purely competitive firm.
   - **d.** producing a standardized product.
4. Which of the following pairs are both "competition-like elements" in monopolistic competition?
   - **a.** Price exceeds MR; standardized product.
   - **b.** Entry is relatively easy; only a normal profit in the long run.
   - **c.** Price equals MC at the profit-maximizing output; economic profits are likely in the long run.
   - **d.** The firms' demand curve is downsloping; differentiated products.

Answers: 1. d; 2. a; 3. a; 4. b

in the industry is producing a specific differentiated product and engaging in a particular amount of advertising. Later we will see how changes in the product and in the amount of advertising modify our conclusions.

The Firm's Demand Curve

Our explanation is based on **Figure 13.1 (Key Graph),** which shows that the demand curve faced by a monopolistically competitive seller is highly, but not perfectly, elastic. It is precisely this feature that distinguishes monopolistic competition from both pure monopoly and pure competition. The monopolistic competitor's demand is more elastic than the demand faced by a pure monopolist because the monopolistically competitive seller has many competitors producing closely substitutable goods. The pure monopolist has no rivals at all. Yet, for two reasons, the monopolistic competitor's demand is not perfectly elastic like that of the pure competitor. First, the monopolistic competitor has fewer rivals; second, its products are differentiated, so they are not perfect substitutes.

The price elasticity of demand faced by the monopolistically competitive firm depends on the number of rivals and the degree of product differentiation. The larger the number of rivals and the weaker the product differentiation, the greater the price elasticity of each seller's demand, that is, the closer monopolistic competition will be to pure competition.

The Short Run: Profit or Loss

In the short run, monopolistically competitive firms maximize profit or minimize loss using exactly the same strategy as pure competitors and monopolists: They produce the level of output at which marginal revenue equals marginal cost (MR = MC). Thus, the monopolistically competitive firm in Figure 13.1a produces output Q_1, where MR = MC. As shown by demand curve D_1, it then can charge price P_1. It realizes an economic profit, shown by the green area [= $(P_1 - A_1) \times Q_1$].

But with less favorable demand or costs, the firm may incur a loss in the short run. We show this possibility in Figure 13.1b, where the firm's best strategy is to minimize its loss. It does so by producing output Q_2 (where MR = MC) and, as determined by demand curve D_2, by charging price P_2. Because price P_2 is less than average total cost A_2, the firm incurs a per-unit loss of $A_2 - P_2$ and a total loss represented as the red area [= $(A_2 - P_2) \times Q_2$].

The Long Run: Only a Normal Profit

In the long run, firms will enter a profitable monopolistically competitive industry and leave an unprofitable one. So a monopolistic competitor will earn only a normal profit in the long run or, in other words, will only break even. (Remember that the cost curves include both explicit and implicit costs, including a normal profit.)

Profits: Firms Enter In the case of short-run profit (Figure 13.1a), economic profits attract new rivals because entry to the industry is relatively easy. As new firms enter, the demand curve faced by the typical firm shifts to the left (falls). Why? Because each firm has a smaller share of total demand and now faces a larger number of close-substitute products. This decline in the firm's demand reduces its economic profit. When entry of new firms has reduced demand to the extent that the demand curve is tangent to the average-total-cost curve at the profit-maximizing output, the firm is just making a normal profit. This situation is shown in Figure 13.1c, where demand is D_3 and the firm's long-run equilibrium output is Q_3. As Figure 13.1c indicates, any greater or lesser output will entail an average total cost that exceeds product price P_3, meaning a loss for the firm. At the tangency point between the demand curve and ATC, total revenue equals total costs. With the economic profit gone, there is no further incentive for additional firms to enter.

Losses: Firms Leave When the industry suffers short-run losses, as in Figure 13.1b, some firms will exit in the long run. Faced with fewer substitute products and blessed with an expanded share of total demand, the surviving firms will see their demand curves shift to the right (rise), as to D_3. Their losses will disappear and give way to normal profits (Figure 13.1c). (For simplicity we have assumed constant costs; shifts in the cost curves as firms enter or leave would complicate our discussion slightly but would not alter our conclusions.)

Complications The representative firm in the monopolistic competition model earns only a normal profit in the long run. That outcome may not always occur, however, in the real world of small firms as opposed to the theoretical model.

- Some firms may achieve sufficient product differentiation such that other firms cannot duplicate them, even over time. One hotel in a major city may have the best location relative to business and tourist activities. Or a firm may have developed a well-known brand name that gives it a slight but very long-lasting advantage over imitators. Such firms may have sufficient monopoly power to realize modest economic profits even in the long run.
- Entry to some industries populated by small firms is not as free in reality as it is in theory. Because of product differentiation, financial barriers to entry are likely to be greater than they would be if the product were standardized. This suggests some monopoly power, with small economic profits continuing even in the long run.

With all things considered, however, the outcome that yields only a normal profit—the long-run equilibrium shown in Figure 13.1c—is a reasonable approximation of reality.

Monopolistic Competition and Efficiency

LO13.3 Explain why monopolistic competition delivers neither productive nor allocative efficiency.

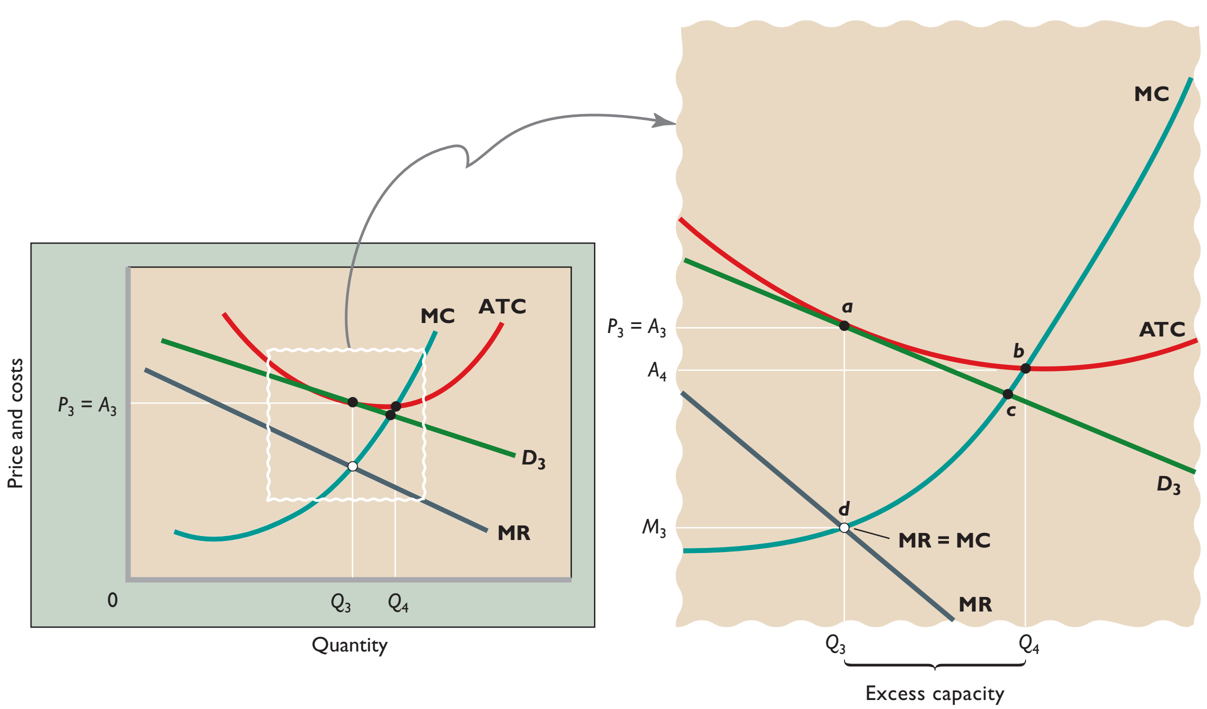
We know from Chapter 11 that economic efficiency requires each firm to produce the amount of output at which P = MC = minimum ATC. The equality of price and minimum average total cost yields *productive efficiency*. The good is being produced in the least costly way, and the price is just sufficient to cover average total cost, including a normal profit. The equality of price and marginal cost yields *allocative efficiency*. The right amount of output is being produced, and thus the right amount of society's scarce resources is being devoted to this specific use.

How efficient is monopolistic competition, as measured against this triple equality? In particular, do monopolistically competitive firms produce the efficient output level associated with P = MC = minimum ATC?

Neither Productive nor Allocative Efficiency

In monopolistic competition, neither productive nor allocative efficiency occurs in long-run equilibrium. Figure 13.2 includes an enlargement of part of Figure 13.1c and clearly shows this. First note that the profit-maximizing price P_3 slightly exceeds the lowest average total cost, A_4. In producing the profit-maximizing output Q_3, the firm's average total cost therefore is slightly higher than optimal from society's perspective—productive efficiency is not achieved. Also note that the profit-maximizing price P_3 exceeds marginal cost (here M_3), meaning that monopolistic competition causes an underallocation of resources. To measure the size of this inefficiency, note that the allocatively optimal amount of output is determined by point *c*, where demand curve D intersects the MC curve. So for all units between Q_3 and the level of output associated with point *c*, marginal benefits exceed marginal costs. Consequently, by producing only Q_3 units, this monopolistic competitor creates an efficiency loss

FIGURE 13.2 The inefficiency of monopolistic competition. In long-run equilibrium a monopolistic competitor achieves neither productive nor allocative efficiency. Productive efficiency is not realized because production occurs where the average total cost A_3 exceeds the minimum average total cost A_4. Allocative efficiency is not achieved because the product price P_3 exceeds the marginal cost M_3. The results are an underallocation of resources as well as an efficiency loss and excess production capacity at every firm in the industry. This firm's efficiency loss is area *acd* and its excess production capacity is $Q_4 - Q_3$.



(deadweight loss) equal in size to area *acd*. The total efficiency loss for the industry as a whole will be the sum of the individual efficiency losses generated by each of the firms in the industry.

Excess Capacity

In monopolistic competition, the gap between the minimum-ATC output and the profit-maximizing output identifies **excess capacity:** plant and equipment that are underused because firms are producing less than the minimum-ATC output. This gap is shown as the distance between Q_4 and Q_3 in Figure 13.2. Note in the figure that the minimum ATC is at point *b*. If each monopolistic competitor could profitably produce at this point on its ATC curve, the lower average total cost would enable a lower price than P_3. More importantly, if each firm produced at *b* rather than at *a*, fewer firms would be needed to produce the industry output. But because monopolistically competitive firms produce at *a* in long-run equilibrium, monopolistically competitive industries are overpopulated with firms, each operating below its optimal capacity. This situation is typified by many kinds of retail establishments. For example, in most cities there is an abundance of small motels and restaurants that operate well below half capacity.

Product Variety

LO13.4 Relate how the ability of monopolistic competition to deliver product differentiation helps to compensate for its failure to deliver economic efficiency.

The situation portrayed in Figures 13.1c and 13.2 is not very satisfying to monopolistic competitors, since it foretells only a normal profit. But the profit-realizing firm of Figure 13.1a need not stand by and watch new competitors eliminate its profit by imitating its product, matching its customer service, and copying its advertising. Each firm has a product that is distinguishable in some way from those of the other producers. So the firm can attempt to stay ahead of competitors and sustain its profit through further product differentiation and better advertising. By developing or improving its product, it may be able to postpone, at least for a while, the outcome of Figure 13.1c.

Although product differentiation and advertising will add to the firm's costs, they can also increase the demand for its product. If demand increases by more than enough to compensate for the added costs, the firm will have improved its profit position. As Figure 13.2 suggests, the firm has little or no prospect of increasing profit by price cutting. So why not engage in nonprice competition?

Benefits of Product Variety

The product variety and product improvement that accompany the drive to maintain economic profit in monopolistic competition are a benefit for society—one that may offset the cost of the inefficiency associated with monopolistic competition. Consumers have a wide diversity of tastes: Some like regular fries, others like curly fries; some like contemporary furniture, others like traditional furniture. If a product is differentiated, then at any time the consumer will be offered a wide range of types, styles, brands, and quality gradations of that product. Compared with pure competition, this provides an advantage to the consumer. The range of choice is widened, and producers more fully meet the wide variation in consumer tastes.

The product improvement promoted by monopolistic competition further differentiates products and expands choices. And a successful product improvement by one firm obligates rivals to imitate or improve on that firm's temporary market advantage or else lose business. So society benefits from better products.

In fact, product differentiation creates a trade-off between consumer choice and productive efficiency. The stronger the product differentiation, the greater is the excess capacity and, therefore, the greater is the productive inefficiency. But the greater the product differentiation, the more likely it is that the firms will satisfy the great diversity of consumer tastes. The greater the excess-capacity problem, the wider the range of consumer choice.

Further Complexity

Finally, the ability to engage in nonprice competition makes the market situation of a monopolistic competitor more complex than Figure 13.1 indicates. That figure assumes a given (unchanging) product and a given level of advertising expenditures. But we know that, in practice, product attributes and advertising are not fixed. The monopolistically competitive firm juggles three factors—price, product, and advertising—in seeking maximum profit. It must determine what variety of product, selling at what price, and supplemented by what level of advertising will result in the greatest profit. This complex situation is not easily expressed in a simple, meaningful economic model. At best, we can say that each possible combination of price, product, and advertising poses a different demand and cost (production cost plus advertising cost) situation for the firm and that one combination yields the maximum profit. In practice, this optimal combination cannot be readily forecast but must be found by trial and error.

QUICK REVIEW 13.1

- Monopolistic competition involves a relatively large number of firms operating in a noncollusive way and producing differentiated products with easy industry entry and exit.
- In the short run, a monopolistic competitor will maximize profit or minimize loss by producing that output at which marginal revenue equals marginal cost.
- In the long run, easy entry and exit of firms cause monopolistic competitors to earn only a normal profit.
- A monopolistic competitor's long-run equilibrium output is such that price exceeds the minimum average total cost (implying that consumers do not get the product at the lowest price attainable) and price exceeds marginal cost (indicating that resources are underallocated to the product).
- The efficiency loss (or deadweight loss) associated with monopolistic competition is greatly muted by the benefits consumers receive from product variety.

Oligopoly

LO13.5 Describe the characteristics of oligopoly.

In terms of competitiveness, the spectrum of market structures reaches from pure competition, to monopolistic competition, to oligopoly, to pure monopoly (review Table 10.1). We now direct our attention to **oligopoly,** a market dominated by a few large producers of a homogeneous or differentiated product. Because of their "fewness," oligopolists have considerable control over their prices, but each must consider the possible reaction of rivals to its own pricing, output, and advertising decisions.

A Few Large Producers

The phrase "a few large producers" is necessarily vague because the market model of oligopoly covers much ground, ranging between pure monopoly, on the one hand, and monopolistic competition, on the other. Oligopoly encompasses the U.S. aluminum industry, in which three huge firms dominate an entire national market, and the situation in which four or five much smaller auto-parts stores enjoy roughly equal shares of the market in a medium-size town. Generally, however, when you hear a term such as "Big Three," "Big Four," or "Big Six," you can be sure it refers to an oligopolistic industry.

Homogeneous or Differentiated Products

An oligopoly may be either a **homogeneous oligopoly** or a **differentiated oligopoly,** depending on whether the firms in the oligopoly produce standardized (homogeneous) or differentiated products. Many industrial products (steel, zinc, copper, aluminum, lead, cement, industrial alcohol) are virtually standardized products that are produced in oligopolies. Alternatively, many consumer goods industries (automobiles, tires, household appliances, electronics equipment, breakfast cereals, cigarettes, and many sporting goods) are differentiated oligopolies. These differentiated oligopolies typically engage in considerable nonprice competition supported by heavy advertising.

Control over Price, but Mutual Interdependence

Because firms are few in oligopolistic industries, each firm is a "price maker"; like the monopolist, it can set its price and output levels to maximize its profit. But unlike the monopolist, which has no rivals, the oligopolist must consider how its rivals will react to any change in its price, output, product characteristics, or advertising. Oligopoly is thus characterized by *strategic behavior* and *mutual interdependence.* By **strategic behavior,** we simply mean self-interested behavior that takes into account the reactions of others. Firms develop and implement price, quality, location, service, and advertising strategies to "grow their business" and expand their profits. But because rivals are few, there is **mutual interdependence:** a situation in which each firm's profit depends not just on its own price and sales strategies but also on those of the other firms in its highly concentrated industry. So oligopolistic firms base their decisions on how they think their rivals will react. Example: In deciding whether to increase the price of its cosmetics, L'Oréal will try to predict the response of the other major producers, such as Clinique. Second example: In deciding on its advertising strategy, Burger King will take into consideration how McDonald's might react.

Entry Barriers

The same barriers to entry that create pure monopoly also contribute to the creation of oligopoly. Economies of scale are important entry barriers in a number of oligopolistic industries, such as the aircraft, rubber, and copper industries. In those industries, three or four firms might each have sufficient sales to achieve economies of scale, but new firms would have such a small market share that they could not do so. They would then be high-cost producers, and as such they could not survive. A closely related barrier is the large expenditure for capital—the cost

CONSIDER THIS . . .

Creative Strategic Behavior

The following story, offered with tongue in cheek, illustrates a localized market that exhibits some characteristics of oligopoly, including strategic behavior.

Tracy Martinez's Native American Arts and Crafts store is located in the center of a small tourist town that borders on a national park. In its early days, Tracy had a minimonopoly. Business was brisk, and prices and profits were high.

To Tracy's annoyance, two "copycat" shops opened adjacent to her store, one on either side of her shop. Worse yet, the competitors named their shops to take advantage of Tracy's advertising. One was "Native Arts and Crafts"; the other, "Indian Arts and Crafts." These new sellers drew business away from Tracy's store, forcing her to lower her prices. The three side-by-side stores in the small, isolated town constituted a localized oligopoly for Native American arts and crafts.

Tracy began to think strategically about ways to boost profit. She decided to distinguish her shop from those on either side by offering a greater mix of high-quality, expensive products and a lesser mix of inexpensive souvenir items. The tactic worked for a while, but the other stores eventually imitated her product mix.

Then, one of the competitors next door escalated the rivalry by hanging up a large sign proclaiming: "We Sell for Less!" Shortly thereafter, the other shop put up a large sign stating: "We Won't Be Undersold!"

Not to be outdone, Tracy painted a colorful sign of her own and hung it above her door. It read: "Main Entrance."

of obtaining necessary plant and equipment—required for entering certain industries. The jet engine, automobile, commercial aircraft, and petroleum-refining industries, for example, are all characterized by very high capital requirements.

The ownership and control of raw materials help explain why oligopoly exists in many mining industries, including gold, silver, and copper. In the computer, chemicals, consumer electronics, and pharmaceutical industries, patents have served as entry barriers. Moreover, oligopolists can preclude the entry of new competitors through preemptive and retaliatory pricing and advertising strategies.

Mergers

Some oligopolies have emerged mainly through the growth of the dominant firms in a given industry (examples: breakfast cereals, chewing gum, candy bars). But for other industries the route to oligopoly has been through mergers (examples: steel, in its early history, and, more recently, airlines, banking, and entertainment). The merging, or combining, of two or more competing firms may substantially increase their market share, and this in turn may allow the new firm to achieve greater economies of scale.

Another motive underlying the "urge to merge" is the desire for monopoly power. The larger firm that results from a merger has greater control over market supply and thus the price of its product. Also, since it is a larger buyer of inputs, it may be able to demand and obtain lower prices (costs) on its production inputs.

Oligopolistic Industries

Previously, we listed the four-firm concentration ratio—the percentage of total industry sales accounted for by the four largest firms—for a number of monopolistically competitive industries (see Table 13.1). Column 2 of Table 13.2 shows the four-firm concentration ratios for 21 oligopolistic industries. For example, the four largest U.S. producers of breakfast cereals make 80 percent of all breakfast cereals produced in the United States.

When the largest four firms in an industry control 40 percent or more of the market (as in Table 13.2), that industry is considered oligopolistic. Using this benchmark, about one-half of all U.S. manufacturing industries are oligopolies.

Although concentration ratios help identify oligopoly, they have four shortcomings.

Localized Markets We have already noted that concentration ratios apply to the nation as a whole, whereas the markets for some products are highly localized because of high transportation costs. Local oligopolies can exist even though national concentration ratios are low.

Interindustry Competition Concentration ratios are based on somewhat arbitrary definitions of industries. In some cases, they disguise significant **interindustry competition**—competition between two products associated with different industries. The high concentration ratio for the copper industry shown in Table 13.2 understates the competition in that industry because aluminum competes with copper in many applications (for example, in the market for long-distance power lines).

TABLE 13.2 **Percentage of Output Produced by Firms in Selected High-Concentration U.S. Manufacturing Industries**

(1) Industry	(2) Percentage of Industry Output* Produced by the Four Largest Firms	(3) Herfindahl Index for the Top 50 Firms
Primary copper	99	ND†
Cane sugar refining	95	ND
Cigarettes	98	ND
Household laundry equipment	98	ND
Household refrigerators and freezers	92	ND
Beer	90	ND
Glass containers	87	2,507
Electronic computers	87	ND
Phosphate fertilizers	83	ND
Aircraft	81	ND

(1) Industry	(2) Percentage of Industry Output* Produced by the Four Largest Firms	(3) Herfindahl Index for the Top 50 Firms
Petrochemicals	80	2,535
Breakfast cereals	80	2,426
Small-arms ammunition	79	2,447
Primary aluminum	77	2,250
Men's slacks and jeans	76	2,015
Electric light bulbs	75	2,258
Tires	73	1,540
Household vacuum cleaners	71	1,519
Alcohol distilleries	70	1,915
Turbines and generators	68	1,937
Motor vehicles	68	1,744

*As measured by value of shipments. Data are for 2007. See **www.census.gov/epcd/www/concentration.html**.

†ND = not disclosed.

Source: Bureau of Census, *Census of Manufacturers, 2007.*

World Trade The data in Table 13.2 only take account of output produced in the United States and may overstate concentration because they do not account for the **import competition** of foreign suppliers. The truck and auto tire industry is a good example. Although Table 13.2 shows that four U.S. firms produce 73 percent of the domestic output of tires, it ignores the fact that a very large portion of the truck and auto tires bought in the United States are imports. Many of the world's largest corporations are foreign, and many of them do business in the United States.

Dominant Firms

The four-firm concentration ratio does not reveal the extent to which one or two firms dominate an industry. Suppose that in industry X one firm produces the entire industry output. In a second industry, Y, four firms compete, each with 25 percent of the market. The concentration ratio is 100 percent for both these industries. But industry X is a pure monopoly, while industry Y is an oligopoly that may be experiencing significant economic rivalry. Most economists would agree that monopoly power (or market power) is substantially greater in industry X than in industry Y, a fact disguised by their identical 100 percent concentration ratios.

The Herfindahl index addresses this problem. Recall that this index is the sum of the squared percentage market shares of all firms in the industry. In equation form:

$$\text{Herfindahl index} = (\%S_1)^2 + (\%S_2)^2 + (\%S_3)^2 + \cdots + (\%S_n)^2$$

where $\%S_1$ is the percentage market share of firm 1, $\%S_2$ is the percentage market share of firm 2, and so on for each firm in the industry. Also remember that by squaring the percentage market shares of all firms in the industry, the Herfindahl index gives much greater weight to larger, and thus more powerful, firms than to smaller ones. In the case of the single-firm industry X, the index would be at its maximum of 100^2, or 10,000, indicating an industry with complete monopoly power. For our supposed four-firm industry Y, the index would be $25^2 + 25^2 + 25^2 + 25^2$, or 2,500, indicating much less market power.

WORKED PROBLEMS

W13.1

Measures of industry competition

The larger the Herfindahl index, the greater the market power within an industry. Note in Table 13.2 that the four-firm concentration ratios for the electric light bulb industry and the tire industry are similar: 75 and 73 percent.

But the Herfindahl index of 2,258 for the electric light bulb industry suggests greater market power than the 1,540 index for the tire industry. Also, contrast the much larger Herfindahl indexes in Table 13.2 with those for the low-concentration industries in Table 13.1.

Oligopoly Behavior: A Game-Theory Overview

LO13.6 Discuss how game theory relates to oligopoly.

Oligopoly pricing behavior has the characteristics of certain games of strategy such as poker, chess, and bridge. The best way to play such a game depends on the way one's opponent plays. Players (and oligopolists) must pattern their actions according to the actions and expected reactions of rivals. The study of how people behave in strategic situations is called **game theory.** A classic example of game theory is called the prisoner's dilemma, in which each of two prisoners confesses to a crime even though they might go free if neither confesses. The logic of this outcome is explained in the nearby Consider This box, which you should read now.

ORIGIN OF THE IDEA

O13.2

Game theory

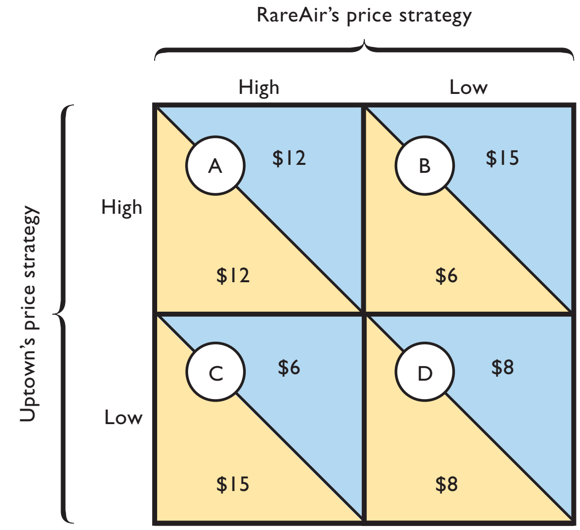
The "confess-confess" outcome of the prisoner's dilemma is conceptually identical to the "low price–low price" outcome in the game shown in Figure 13.3. In Figure 13.3 we assume that a duopoly, or two-firm oligopoly, is producing athletic shoes. Each of the two firms—let's call them RareAir and Uptown—has a choice of two pricing strategies: price high or price low. The profit each firm earns will depend on the strategy it chooses *and* the strategy its rival chooses.

There are four possible combinations of strategies for the two firms, and a lettered cell in Figure 13.3 represents each combination. For example, cell C represents a low-price strategy for Uptown along with a high-price strategy for RareAir. Figure 13.3 is called a *payoff matrix* because each cell shows the payoff (profit) to each firm that would result from each combination of strategies. Cell C shows that if Uptown adopts a low-price strategy and RareAir a high-price strategy, then Uptown will earn $15 million (yellow portion) and RareAir will earn $6 million (blue portion).

Mutual Interdependence Revisited

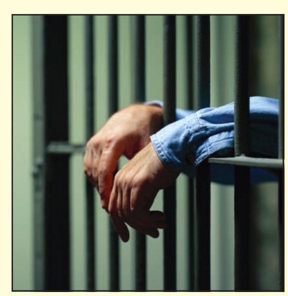
The data in Figure 13.3 are hypothetical, but their relationships are typical of real situations. Recall that oligopolistic firms can increase their profits, and influence their rivals' profits, by changing their pricing strategies. Each firm's profit depends on its own pricing strategy and that of its

FIGURE 13.3 **Profit payoff (in millions) for a two-firm oligopoly.** Each firm has two possible pricing strategies. RareAir's strategies are shown in the top margin, and Uptown's in the left margin. Each lettered cell of this four-cell payoff matrix represents one combination of a RareAir strategy and an Uptown strategy and shows the profit that combination would earn for each. Assuming no collusion, the outcome of this game is Cell D, with both parties using low-price strategies and earning $8 million of profits.



CONSIDER THIS . . .

The Prisoner's Dilemma



One of the classic illustrations of game theory is the prisoner's dilemma game in which two people—let's call them Betty and Al—have committed a diamond heist and are being detained by the police as prime suspects. Unbeknownst to the two, the evidence against them is weak, so the best hope the police have for getting a conviction is if one or both of the thieves confess to the crime. The police place Betty and Al in separate holding cells and offer each the same deal: Confess to the crime and receive a lighter prison sentence.

Each detainee therefore faces a dilemma. If Betty remains silent and Al confesses, Betty will end up with a long prison sentence. If Betty confesses and Al says nothing, Al will receive a long prison sentence. What happens? Fearful that the other person will confess, both confess, even though they each would be better off saying nothing.

rivals. This mutual interdependence of oligopolists is the most obvious point demonstrated by Figure 13.3. If Uptown adopts a high-price strategy, its profit will be $12 million provided that RareAir also employs a high-price strategy (cell A). But if RareAir uses a low-price strategy against Uptown's high-price strategy (cell B), RareAir will increase its market share and boost its profit from $12 to $15 million. RareAir's higher profit will come at the expense of Uptown, whose profit will fall from $12 million to $6 million. Uptown's high-price strategy is a good strategy only if RareAir also employs a high-price strategy.

Collusion

Figure 13.3 also suggests that oligopolists often can benefit from **collusion**—that is, cooperation with rivals. To see the benefits of collusion, first suppose that both firms in Figure 13.3 are acting independently and following high-price strategies. Each realizes a $12 million profit (cell A).

Note that either RareAir or Uptown could increase its profit by switching to a low-price strategy (cell B or C). The low-price firm would increase its profit to $15 million and the high-price firm's profit would fall to $6 million. The high-price firm would be better off if it, too, adopted a low-price policy. Doing so would increase its profit from $6 million to $8 million (cell D). The effect of all this independent strategy shifting would be the reduction of both firms' profits from $12 million (cell A) to $8 million (cell D).

In real situations, too, independent action by oligopolists may lead to mutually "competitive" low-price strategies: Independent oligopolists compete with respect to price, and this leads to lower prices and lower profits. This outcome is clearly beneficial to consumers but not to the oligopolists, whose profits decrease.

How could oligopolists avoid the low-profit outcome of cell D? The answer is that they could collude, rather than establish prices competitively or independently. In our example, the two firms could agree to establish and maintain a high-price policy. So each firm will increase its profit from $8 million (cell D) to $12 million (cell A).

Incentive to Cheat

The payoff matrix also explains why an oligopolist might be strongly tempted to cheat on a collusive agreement. Suppose Uptown and RareAir agree to maintain high-price policies, with each earning $12 million in profit (cell A). Both are tempted to cheat on this collusive pricing agreement because either firm can increase its profit to $15 million by lowering its price.

For instance, if Uptown secretly cheats and sells at the low price while RareAir keeps on charging the high price, the payoff would move from cell A to cell C so that Uptown's profit would rise to $15 million while RareAir's profit would fall to $6 million. On the other hand, if RareAir cheats and sets a low price while Uptown keeps the agreement and charges the high price, the payoff matrix would move from cell A to cell B so that RareAir would get $15 million while Uptown would get only $6 million.

As you can see, cheating is both very lucrative to the cheater as well as very costly to the firm that gets cheated on. As a result, both firms will probably cheat so that the game will settle back to cell D, with each firm using its low-price strategy. (The Consider This box on the prisoner's dilemma is highly relevant and we urge you to read it now. Also, the appendix to this chapter provides several additional applications of game theory.)

QUICK REVIEW 13.2

- An oligopoly is made up of relatively few firms producing either homogeneous or differentiated products; these firms are mutually interdependent.
- Barriers to entry such as scale economies, control of patents or strategic resources, or the ability to engage in retaliatory pricing characterize oligopolies. Oligopolies may result from internal growth of firms, mergers, or both.
- The four-firm concentration ratio shows the percentage of an industry's sales accounted for by its four largest firms; the Herfindahl index measures the degree of market power in an industry by summing the squares of the percentage market shares held by the individual firms in the industry.
- Game theory reveals that (a) oligopolies are mutually interdependent in their pricing policies; (b) collusion enhances oligopoly profits; and (c) there is a temptation for oligopolists to cheat on a collusive agreement.

Three Oligopoly Models

LO13.7 Explain the three main models of oligopoly pricing and output: kinked-demand theory, collusive pricing, and price leadership.

To gain further insight into oligopolistic pricing and output behavior, we will examine three distinct pricing models: (1) the kinked-demand curve, (2) collusive pricing, and (3) price leadership.

Why not a single model, as in our discussions of the other market structures? There are two reasons:

- ***Diversity of oligopolies*** Oligopoly encompasses a greater range and diversity of market situations than do other market structures. It includes the *tight* oligopoly, in which two or three firms dominate an entire market, and the *loose* oligopoly, in which six or seven firms share, say, 70 or 80 percent of a market while a "competitive fringe" of firms shares the remainder. It includes both differentiated and standardized products. It includes cases in which firms act in collusion and those in which they act independently. It embodies situations in which barriers to entry are very strong and situations in which they are not quite so strong. In short, the diversity of oligopoly does not allow us to explain all oligopolistic behaviors with a single market model.

- ***Complications of interdependence*** The mutual interdependence of oligopolistic firms complicates matters significantly. Because firms cannot predict the reactions of their rivals with certainty, they cannot estimate their own demand and marginal-revenue data. Without such data, firms cannot determine their profit-maximizing price and output, even in theory, as we will see.

Despite these analytical difficulties, two interrelated characteristics of oligopolistic pricing have been observed. First, if the macroeconomy is generally stable, oligopolistic prices are typically inflexible (or "rigid" or "sticky"). Prices change less frequently under oligopoly than under pure competition, monopolistic competition, and, in some instances, pure monopoly. Second, when oligopolistic prices do change, firms are likely to change their prices together, suggesting that there is a tendency to act in concert, or collusively, in setting and changing prices (as we mentioned in the preceding section). The diversity of oligopolies and the presence of mutual interdependence are reflected in the models that follow.

Kinked-Demand Theory: Noncollusive Oligopoly

Imagine an oligopolistic industry made up of three hypothetical firms (Arch, King, and Dave's), each having about one-third of the total market for a differentiated product. Assume that the firms are "independent," meaning that they do not engage in collusive price practices. Assume, too, that the going price for Arch's product is P_0 and its current sales are Q_0, as shown in **Figure 13.4a (Key Graph).**

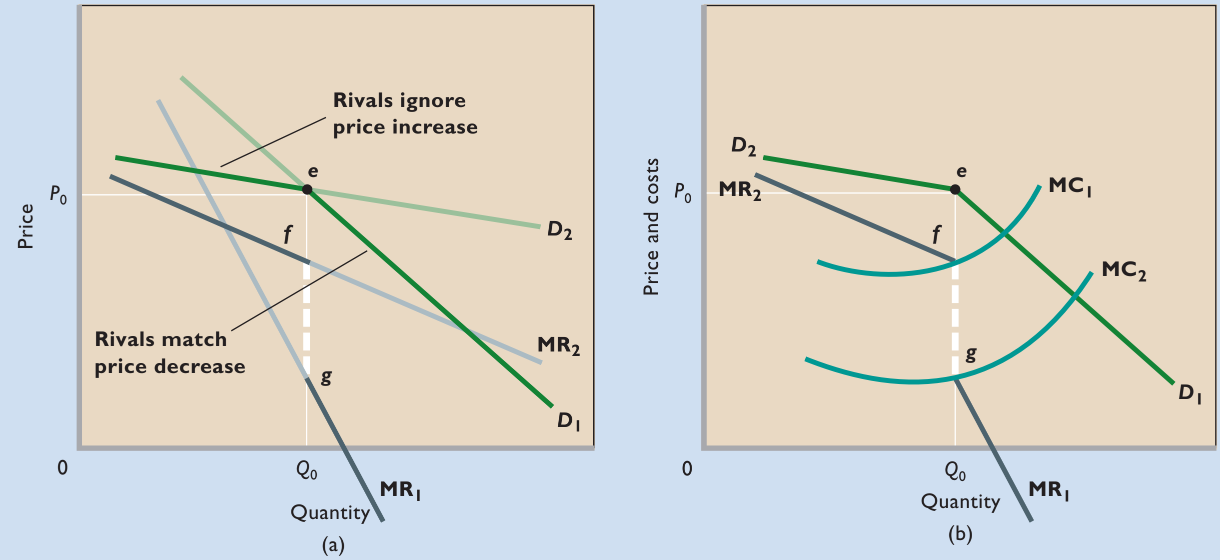
Now the question is, "What does the firm's demand curve look like?" Mutual interdependence and the uncertainty about rivals' reactions make this question hard to answer. The location and shape of an oligopolist's demand curve depend on how the firm's rivals will react to a price change introduced by Arch. There are two plausible assumptions about the reactions of Arch's rivals:

- ***Match price changes*** One possibility is that King and Dave's will exactly match any price change initiated by Arch. In this case, Arch's demand and marginal-revenue curves will look like the straight lines labeled D_1 and MR_1 in Figure 13.4a. Why are they so steep? Reason: If Arch cuts its price, its sales will increase only modestly because its two rivals will also cut their prices to prevent Arch from gaining an advantage over them. The small increase in sales that Arch (and its two rivals) will realize is at the expense of other industries; Arch will gain no sales from King and Dave's. In a similar fashion, if Arch raises its price, its sales will fall only modestly because King and Dave's will match its price increase. The industry will lose sales to other industries, but Arch will lose no customers to King and Dave's.
- ***Ignore price changes*** The other possibility is that King and Dave's will ignore any price change by Arch. In this case, the demand and marginal-revenue curves faced by Arch will resemble the straight lines D_2 and MR_2 in Figure 13.4a. Demand in this case is considerably more elastic than it was under the previous assumption. The reasons are clear: If Arch lowers its price and its rivals do not, Arch will gain sales significantly at the expense of its two rivals because it will be underselling them. Conversely, if Arch raises its price and its rivals do not, Arch will lose many customers to King and Dave's, which will be underselling it. Because of product differentiation, however, Arch's sales will not fall to zero when it raises its price; some of Arch's customers will pay the higher price because they have a strong preference for Arch's product. Nevertheless, Arch's demand curve will be much more elastic when its rivals ignore price changes than when they match them.

A Combined Strategy Now, which is the most logical assumption for Arch to make about how its rivals will react to any price change it might initiate? The answer is, "It depends on the direction of the price change." Common sense and observation of oligopolistic industries suggest that a firm's rivals will match price declines below P_0 as they act to prevent the price cutter from taking their customers. But they will ignore price increases above P_0 because the rivals of the price-increasing firm stand to gain the business lost by the price booster. In other words, the dark-green left-hand segment of the "rivals ignore" demand curve D_2 in Figure 13.4a seems relevant for price increases, and the dark-green right-hand segment of the "rivals

KEY GRAPH

FIGURE 13.4 The kinked-demand curve. (a) The slope of a noncollusive oligopolist's demand and marginal-revenue curves depends on whether its rivals match (straight lines D_1 and MR_1) or ignore (straight lines D_2 and MR_2) any price changes that it may initiate from the current price P_0. (b) In all likelihood an oligopolist's rivals will ignore a price increase but follow a price cut. This causes the oligopolist's demand curve to be kinked (D_2eD_1) and the marginal-revenue curve to have a vertical break, or gap (*fg*). Because any shift in marginal costs between MC_1 and MC_2 will cut the vertical (dashed) segment of the marginal-revenue curve, no change in either price P_0 or output Q_0 will result from such a shift.



QUICK QUIZ FOR FIGURE 13.4

1. Suppose Q_0 in this figure represents annual sales of 5 million units for this firm. The other two firms in this three-firm industry sell 3 million and 2 million units, respectively. The Herfindahl index for this industry is:
   a. 100 percent.
   b. 400.
   c. 10.
   d. 3,800.
2. The D_2e segment of the demand curve D_2eD_1 in graph (b) implies that:
   a. this firm's total revenue will fall if it increases its price above P_0.
   b. other firms will match a price increase above P_0.
   c. the firm's relevant marginal-revenue curve will be MR_1 for price increases above P_0.
   d. the product in this industry is necessarily standardized.
3. By matching a price cut, this firm's rivals can:
   a. increase their market shares.
   b. increase their marginal revenues.
   c. maintain their market shares.
   d. lower their total costs.
4. A shift of the marginal-cost curve from MC_2 to MC_1 in graph (b) would:
   a. increase the "going price" above P_0.
   b. leave price at P_0 but reduce this firm's total profit.
   c. leave price at P_0 but reduce this firm's total revenue.
   d. make this firm's demand curve more elastic.

Answers: 1. d; 2. a; 3. c; 4. b

match" demand curve D_1 seems relevant for price cuts. It is therefore reasonable to assume that the noncollusive oligopolist faces the **kinked-demand curve** D_2eD_1, as shown in Figure 13.4b. Demand is highly elastic above the going price P_0 but much less elastic or even inelastic below that price.

Note also that if rivals match a price cut but ignore an increase, the marginal-revenue curve of the oligopolist will also have an odd shape. It, too, will be made up of two segments: the dark gray left-hand part of marginal-revenue curve MR_2 in Figure 13.4a and the dark gray right-hand part of marginal-revenue curve MR_1. Because of the sharp difference in elasticity of demand above and below the going price, there is a gap, or what we can simply treat as a vertical segment, in the marginal-revenue curve. We show this gap as the dashed segment in the combined marginal-revenue curve MR_2fgMR_1 in Figure 13.4b.

Price Inflexibility This analysis helps explain why prices are generally stable in noncollusive oligopolistic industries. There are both demand and cost reasons.

On the demand side, the kinked-demand curve gives each oligopolist reason to believe that any change in price will be for the worse. If it raises its price, many of its customers will desert it. If it lowers its price, its sales at best will increase very modestly since rivals will match the lower price. Even if a price cut increases the oligopolist's total revenue somewhat, its costs may increase by a greater amount, depending on demand elasticity. For instance, if its demand is inelastic to the right of Q_0, as it may well be, then the firm's profit will surely fall. A price decrease in the inelastic region lowers the firm's total revenue, and the production of a larger output increases its total costs.

On the cost side, the broken marginal-revenue curve suggests that even if an oligopolist's costs change substantially, the firm may have no reason to change its price. In particular, all positions of the marginal-cost curve between MC_1 and MC_2 in Figure 13.4b will result in the firm's deciding on exactly the same price and output. For all those positions, MR equals MC at output Q_0; at that output, it will charge price P_0.

Criticisms of the Model The kinked-demand analysis has two shortcomings. First, it does not explain how the going price gets to be at P_0 in Figure 13.4 in the first place. It only helps explain why oligopolists tend to stick with an existing price. The kinked-demand curve explains price inflexibility but not price itself.

Second, when the macroeconomy is unstable, oligopoly prices are not as rigid as the kinked-demand theory implies. During inflationary periods, many oligopolists have raised their prices often and substantially. And during downturns (recessions), some oligopolists have cut prices. In some instances these price reductions have set off a **price war:** successive and continuous rounds of price cuts by rivals as they attempt to maintain their market shares.

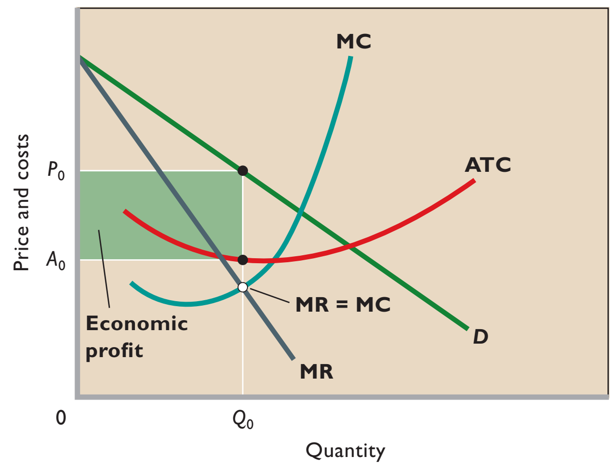
Cartels and Other Collusion

Our game-theory model demonstrated that oligopolists might benefit from collusion. We can say that collusion occurs whenever firms in an industry reach an agreement to fix prices, divide up the market, or otherwise restrict competition among themselves. The disadvantages and uncertainties of noncollusive, kinked-demand oligopolies are obvious. There is always the danger of a price war breaking out, especially during a general business recession. Then each firm finds that, because of unsold goods and excess capacity, it can reduce per-unit costs by increasing market share. Then, too, a new firm may surmount entry barriers and initiate aggressive price cutting to gain a foothold in the market. In addition, the kinked-demand curve's tendency toward rigid prices may adversely affect profits if general inflationary pressures increase costs. However, by controlling price through collusion, oligopolists may be able to reduce uncertainty, increase profits, and perhaps even prohibit the entry of new rivals.

Price and Output Assume once again that there are three hypothetical oligopolistic firms (Gypsum, Sheetrock, and GSR) producing, in this instance, gypsum drywall panels for finishing interior walls. All three firms produce a homogeneous product and have identical cost curves. Each firm's demand curve is indeterminate unless we know how its rivals will react to any price change. Therefore, we suppose each firm assumes that its two rivals will match either a price cut or a price increase. In other words, each firm has a demand curve like the straight line D_1 in Figure 13.4a. And since they have identical cost data, and the same demand and thus marginal-revenue data, we can say that Figure 13.5 represents the position of each of our three oligopolistic firms.

What price and output combination should, say, Gypsum select? If Gypsum were a pure monopolist, the answer would be clear: Establish output at Q_0, where

FIGURE 13.5 Collusion and the tendency toward joint-profit maximization. If oligopolistic firms face identical or highly similar demand and cost conditions, they may collude to limit their joint output and to set a single, common price. Thus each firm acts as if it were a pure monopolist, setting output at Q_0 and charging price P_0. This price and output combination maximizes each oligopolist's profit (green area) and thus the combined or joint profit of the colluding firms.



marginal revenue equals marginal cost, charge the corresponding price P_0, and enjoy the maximum profit attainable. However, Gypsum does have two rivals selling identical products, and if Gypsum's assumption that its rivals will match its price of P_0 proves to be incorrect, the consequences could be disastrous for Gypsum. Specifically, if Sheetrock and GSR actually charge prices below P_0, then Gypsum's demand curve D will shift sharply to the left as its potential customers turn to its rivals, which are now selling the same product at a lower price. Of course, Gypsum can retaliate by cutting its price too, but this will move all three firms down their demand curves, lowering their profits. It may even drive them to a point where average total cost exceeds price and losses are incurred.

So the question becomes, "Will Sheetrock and GSR want to charge a price below P_0?" Under our assumptions, and recognizing that Gypsum has little choice except to match any price they may set below P_0, the answer is no. Faced with the same demand and cost circumstances, Sheetrock and GSR will find it in their interest to produce Q_0 and charge P_0. This is a curious situation; each firm finds it most profitable to charge the same price, P_0, but only if its rivals actually do so! How can the three firms ensure the price P_0 and quantity Q_0 solution in which each is keenly interested? How can they avoid the less profitable outcomes associated with either higher or lower prices?

The answer is evident: They can collude. They can get together, talk it over, and agree to charge the same price, P_0, and thereby enjoy the maximum profit available [$= (P_0 - A_0) \times Q_0$ units]. In addition to reducing the possibility of price wars, this will give each firm the maximum profit. (But it will also subject them to antitrust prosecution if they are caught!) For society, the result will be the same as would occur if the industry were a pure monopoly composed of three identical plants.

Overt Collusion: The OPEC Cartel Collusion may assume a variety of forms. The most comprehensive form of collusion is the **cartel,** a group of producers that typically creates a formal written agreement specifying how much each member will produce and charge. Output must be controlled—the market must be divided up—in order to maintain the agreed-upon price. The collusion is overt, or open to view.

Undoubtedly the most significant international cartel is the Organization of Petroleum Exporting Countries (OPEC), comprising 12 oil-producing nations (see Global Perspective 13.1). OPEC produces 43 percent of the world's oil and supplies 45 percent of all oil traded internationally.

GLOBAL PERSPECTIVE 13.1

The 12 OPEC Nations, Daily Oil Production, October 2012

The OPEC nations produce about 43 percent of the world's oil and about 45 percent of the oil sold in world markets.

OPEC Country	Barrels of Oil
Saudi Arabia	9,311,000
Iran	3,576,000
Venezuela	2,881,000
Kuwait	2,659,000
Iraq	2,653,000
UAE	2,565,000
Nigeria	1,975,000
Angola	1,618,000
Algeria	1,162,000
Qatar	734,000
Ecuador	500,000
Libya	489,000

Source: OPEC, **www.opec.org**.

OPEC has in some cases been able to drastically alter oil prices by increasing or decreasing supply. In 1973, for instance, it caused the price of oil to more than triple by getting its members to restrict output. And again, in the late 1990s it caused oil prices to rise from \$11 per barrel to \$34 per barrel over a 15-month period.

That being said, it should be kept in mind that most increases in the price of oil are not caused by OPEC. Between 2005 and 2008, for example, oil prices went from \$40 per barrel to \$140 per barrel due to rapidly rising demand from China and supply uncertainties related to armed conflict in the Middle East. But as the recession that began in December 2007 took hold, demand slumped and oil prices collapsed back down to about \$40 per barrel. OPEC was largely a nonfactor in this rise and fall in the price of oil. But in those cases where OPEC can effectively enforce its production agreements, there is little doubt that it can hold the price of oil substantially above the marginal cost of production.

Covert Collusion: Examples Cartels are illegal in the United States, and hence any collusion that exists is covert or secret. Yet there are numerous examples, as shown by evidence from antitrust (antimonopoly) cases. In 2011, U.S.-based Whirlpool, Japan-headquartered Panasonic, the Danish firm Danfoss, and the Italian company Appliance Components were fined over $200 million for attempting to run an international cartel that could rig the worldwide prices of refrigerator compressors. In 2012, several Japanese autoparts makers pleaded guilty to rigging the bids that they submitted to a major carmaker. The conspirators employed measures to keep their conduct secret, including using code names and instructing participants to destroy evidence of collusion.

In many other instances collusion is much subtler. Unwritten, informal understandings (historically called "gentlemen's agreements") are frequently made at cocktail parties, on golf courses, through phone calls, or at trade association meetings. In such agreements, executives reach verbal or even tacit (unspoken) understandings on product price, leaving market shares to be decided by nonprice competition. Although these agreements, too, violate antitrust laws—and can result in severe personal and corporate penalties—the elusive character of informal understandings makes them more difficult to detect.

Obstacles to Collusion Normally, cartels and similar collusive arrangements are difficult to establish and maintain. Here are several barriers to collusion:

Demand and Cost Differences When oligopolists face different costs and demand curves, it is difficult for them to agree on a price. This is particularly the case in industries where products are differentiated and change frequently. Even with highly standardized products, firms usually have somewhat different market shares and operate with differing degrees of productive efficiency. Thus it is unlikely that even homogeneous oligopolists would have the same demand and cost curves.

In either case, differences in costs and demand mean that the profit-maximizing price will differ among firms; no single price will be readily acceptable to all, as we assumed was true in Figure 13.5. So price collusion depends on compromises and concessions that are not always easy to obtain and hence act as an obstacle to collusion.

Number of Firms Other things equal, the larger the number of firms, the more difficult it is to create a cartel or some other form of price collusion. Agreement on price by three or four producers that control an entire market may be relatively easy to accomplish. But such agreement is more difficult to achieve where there are, say, 10 firms, each with roughly 10 percent of the market, or where the Big Three have 70 percent of the market while a competitive fringe of 8 or 10 smaller firms battles for the remainder.

Cheating As the game-theory model makes clear, collusive oligopolists are tempted to engage in secret price cutting to increase sales and profit. The difficulty with such cheating is that buyers who are paying a high price for a product may become aware of the lower-priced sales and demand similar treatment. Or buyers receiving a price concession from one producer may use the concession as a wedge to get even larger price concessions from a rival producer. Buyers' attempts to play producers against one another may precipitate price wars among the producers. Although secret price concessions are potentially profitable, they threaten collusive oligopolies over time. Collusion is more likely to succeed when cheating is easy to detect and punish. Then the conspirators are less likely to cheat on the price agreement.

Recession Long-lasting recession usually serves as an enemy of collusion because slumping markets increase average total cost. In technical terms, as the oligopolists' demand and marginal-revenue curves shift to the left in Figure 13.5 in response to a recession, each firm moves leftward and upward to a higher operating point on its average-total-cost curve. Firms find they have substantial excess production capacity, sales are down, unit costs are up, and profits are being squeezed. Under such conditions, businesses may feel they can avoid serious profit reductions (or even losses) by cutting price and thus gaining sales at the expense of rivals.

Potential Entry The greater prices and profits that result from collusion may attract new entrants, including foreign firms. Since that would increase market supply and reduce prices and profits, successful collusion requires that colluding oligopolists block the entry of new producers.

Legal Obstacles: Antitrust Law U.S. antitrust laws prohibit cartels and price-fixing collusion. So less obvious means of price control have evolved in the United States.

Price Leadership Model

Price leadership entails a type of implicit understanding by which oligopolists can coordinate prices without engaging in outright collusion based on formal agreements and secret meetings. Rather, a practice evolves whereby the "dominant firm"—usually the largest or most efficient in the industry—initiates price changes and all other firms more or less automatically follow the leader. Many industries, including farm machinery, cement, copper, newsprint, glass containers, steel, beer, fertilizer, cigarettes, and tin, are practicing, or have in the recent past practiced, price leadership.

Leadership Tactics An examination of price leadership in a variety of industries suggests that the price leader is likely to observe the following tactics.

Infrequent Price Changes Because price changes always carry the risk that rivals will not follow the lead, price adjustments are made only infrequently. The price leader does not respond to minuscule day-to-day changes in costs and demand. Price is changed only when cost and demand conditions have been altered significantly and on an industrywide basis as the result of, for example, industrywide wage increases, an increase in excise taxes, or an increase in the price of some basic input such as energy. In the automobile industry, price adjustments traditionally have been made when new models are introduced each fall.

Communications The price leader often communicates impending price adjustments to the industry through speeches by major executives, trade publication interviews, or press releases. By publicizing "the need to raise prices," the price leader seeks agreement among its competitors regarding the actual increase.

Limit Pricing The price leader does not always choose the price that maximizes short-run profits for the industry because the industry may want to discourage new firms from entering. If the cost advantages (economies of scale) of existing firms are a major barrier to entry, new entrants could surmount that barrier if the price leader and the other firms set product price high enough. New firms that are relatively inefficient because of their small size might survive and grow if the industry sets price very high. So, in order to discourage new competitors and to maintain the current oligopolistic structure of the industry, the price leader may keep price below the short-run profit-maximizing level. The strategy of establishing a price that blocks the entry of new firms is called *limit pricing*.

Breakdowns in Price Leadership: Price Wars Price leadership in oligopoly occasionally breaks down, at least temporarily, and sometimes results in a price war. An example of price leadership temporarily breaking down occurred in the breakfast cereal industry, in which Kellogg traditionally had been the price leader. General Mills countered Kellogg's leadership in 1995 by reducing the prices of its cereals by 11 percent. In 1996 Post responded with a 20 percent price cut, which Kellogg then followed. Not to be outdone, Post reduced its prices by another 11 percent.

As another example, in October 2009 with the Christmas shopping season just getting underway, Walmart cut its price on 10 highly anticipated new books to just $10 each. Within hours, Amazon.com matched the price cut. Walmart then retaliated by cutting its price for the books to just $9 each. Amazon.com matched that reduction—at which point Walmart went to $8.99! Then, out of nowhere, Target jumped in at $8.98, a price that Amazon.com and Walmart immediately matched. And that is where the price finally came to rest—at a level so low that each company was losing money on each book it sold.

Most price wars eventually run their course. After a period of low or negative profits, they again yield price leadership to one of the industry's leading firms. That firm then begins to raise prices, and the other firms willingly follow suit.

QUICK REVIEW 13.3

- In the kinked-demand theory of oligopoly, price is relatively inflexible because a firm contemplating a price change assumes that its rivals will follow a price cut and ignore a price increase.
- Cartels agree on production limits and set a common price to maximize the joint profit of their members as if each were a subsidiary of a single pure monopoly.
- Collusion among oligopolists is difficult because of (a) demand and cost differences among sellers, (b) the complexity of output coordination among producers, (c) the potential for cheating, (d) a tendency for agreements to break down during recessions, (e) the potential entry of new firms, and (f) antitrust laws.
- Price leadership involves an informal understanding among oligopolists to match any price change initiated by a designated firm (often the industry's dominant firm).

Oligopoly and Advertising

LO13.8 Contrast the potential positive and negative effects of advertising.

We have noted that oligopolists would rather not compete on the basis of price and may become involved in price collusion. Nonetheless, each firm's share of the total market is typically determined through product development and advertising, for two reasons:

- Product development and advertising campaigns are less easily duplicated than price cuts. Price cuts can be quickly and easily matched by a firm's rivals to cancel any potential gain in sales derived from that strategy. Product improvements and successful advertising, however, can produce more permanent gains in market share because they cannot be duplicated as quickly and completely as price reductions.

- Oligopolists have sufficient financial resources to engage in product development and advertising. For most oligopolists, the economic profits earned in the past can help finance current advertising and product development.

Product development (or, more broadly, "research and development") is the subject of Web Chapter 13, so we will confine our present discussion to advertising. In 2011, firms spent an estimated $103 billion on advertising in the United States and $498 billion worldwide. Advertising is prevalent in both monopolistic competition and oligopoly. Table 13.3 lists the 10 leading U.S. advertisers in 2011.

Advertising may affect prices, competition, and efficiency both positively and negatively, depending on the circumstances. While our focus here is on advertising by oligopolists, the analysis is equally applicable to advertising by monopolistic competitors.

Positive Effects of Advertising

In order to make rational (efficient) decisions, consumers need information about product characteristics and prices. Media advertising may be a low-cost means for consumers to obtain that information. Suppose you are in the market for a high-quality camera that is not advertised or promoted in newspapers, in magazines, or on the Internet. To make a rational choice, you may have to spend several days visiting stores to determine the availability, prices, and features of various brands. This search entails both direct costs (gasoline, parking fees) and indirect costs (the value of your time). By providing information about the available options, advertising and Internet promotion reduce your search time and minimize these direct and indirect costs.

By providing information about the various competing goods that are available, advertising diminishes monopoly power. In fact, advertising is frequently associated with the introduction of new products designed to compete with existing brands. Could Toyota and Honda have so strongly challenged U.S. auto producers without advertising? Could FedEx have sliced market share away from UPS and the U.S. Postal Service without advertising?

Viewed this way, advertising is an efficiency-enhancing activity. It is a relatively inexpensive means of providing useful information to consumers and thus lowering their search costs. By enhancing competition, advertising results in greater economic efficiency. By facilitating the introduction of new products, advertising speeds up technological progress. By increasing sales and output, advertising can reduce long-run average total cost by enabling firms to obtain economies of scale.

TABLE 13.3 **The Largest U.S. Advertisers, 2011**

Company	Advertising Spending Millions of $
Procter & Gamble	$4,971
General Motors	3,055
Verizon	2,523
Comcast	2,465
AT&T	2,359
JPMorgan Chase	2,351
Ford Motor	2,141
American Express	2,125
L'Oréal	2,124
Walt Disney	2,112

Source: Advertising Age, **www.adage.com**. Copyright *Global AdView Pulse lite,* Copyright The Nielsen Company, 2012. Crain Communications. 69284-36mpf.

Potential Negative Effects of Advertising

Not all the effects of advertising are positive, of course. Much advertising is designed simply to manipulate or persuade consumers—that is, to alter their preferences in favor of the advertiser's product. A television commercial that indicates that a popular personality drinks a particular brand of soft drink—and therefore that you should too—conveys little or no information to consumers about price or quality. In addition, advertising is sometimes based on misleading and extravagant claims that confuse consumers rather than enlighten them. Indeed, in some cases advertising may well persuade consumers to pay high prices for much-acclaimed but inferior products, forgoing better but unadvertised products selling at lower prices. Example: *Consumer Reports* has found that heavily advertised premium motor oils provide no better engine performance and longevity than do cheaper brands.

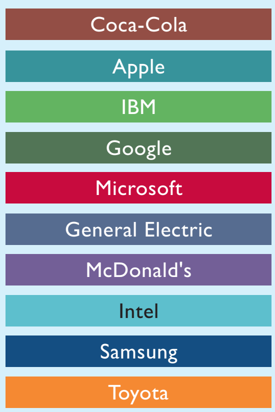
Firms often establish substantial brand-name loyalty and thus achieve monopoly power via their advertising (see Global Perspective 13.2). As a consequence, they are able to increase their sales, expand their market shares, and enjoy greater profits. Larger profits permit still more advertising and further enlargement of the firm's market share and profit. In time, consumers may lose the advantages of competitive markets and face the disadvantages of monopolized markets. Moreover, new entrants to the

GLOBAL PERSPECTIVE 13.2

The World's Top 10 Brand Names, 2012

Here are the world's top 10 brands, based on four criteria: the brand's market share within its category, the brand's world appeal across age groups and nationalities, the loyalty of customers to the brand, and the ability of the brand to "stretch" to products beyond the original product.

World's Top 10 Brands



Source: 100 Best Global Brands, 2012. Used with permission of Interbrand, **www.interbrand.com**.

industry need to incur large advertising costs in order to establish their products in the marketplace; thus, advertising costs may be a barrier to entry.

Advertising can also be self-canceling. The advertising campaign of one fast-food hamburger chain may be offset by equally costly campaigns waged by rivals, so each firm's demand actually remains unchanged. Few, if any, extra burgers will be purchased and each firm's market share will stay the same. But because of the advertising, all firms will experience higher costs and either their profits will fall or, through successful price leadership, their product prices will rise.

When advertising either leads to increased monopoly power or is self-canceling, economic inefficiency results.

Oligopoly and Efficiency

LO13.9 Discuss the efficiency of oligopoly from society's standpoint and whether it is more or less efficient than monopoly.

Is oligopoly, then, an efficient market structure from society's standpoint? How do the price and output decisions of the oligopolist measure up to the triple equality $P = MC =$ minimum ATC that occurs in pure competition?

Productive and Allocative Efficiency

Many economists believe that the outcome of some oligopolistic markets is approximately as shown in Figure 13.5. This view is bolstered by evidence that many oligopolists sustain sizable economic profits year after year. In that case, the oligopolist's production occurs where price exceeds marginal cost and average total cost. Moreover, production is below the output at which average total cost is minimized. In this view, neither productive efficiency ($P =$ minimum ATC) nor allocative efficiency ($P =$ MC) is likely to occur under oligopoly.

A few observers assert that oligopoly is actually less desirable than pure monopoly because government usually regulates pure monopoly in the United States to guard against abuses of monopoly power. Informal collusion among oligopolists may yield price and output results similar to those under pure monopoly yet give the outward appearance of competition involving independent firms.

Qualifications

We should note, however, three qualifications to this view:

- ***Increased foreign competition*** In recent decades foreign competition has increased rivalry in a number of oligopolistic industries—steel, automobiles, video games, electric shavers, outboard motors, and copy machines, for example. This has helped to break down such cozy arrangements as price leadership and to stimulate much more competitive pricing.
- ***Limit pricing*** Recall that some oligopolists may purposely keep prices below the short-run profit-maximizing level in order to bolster entry barriers. In essence, consumers and society may get some of the benefits of competition—prices closer to marginal cost and minimum average total cost—even without the competition that free entry would provide.
- ***Technological advance*** Over time, oligopolistic industries may foster more rapid product development and greater improvement of production techniques than would be possible if they were purely competitive. Oligopolists have large economic profits from which they can fund expensive research and development (R&D). Moreover, the existence of barriers to entry may give the oligopolist some assurance that it will reap the rewards of successful R&D. Thus, the short-run economic inefficiencies of oligopolists may be partly or wholly offset by the oligopolists' contributions to better products, lower prices, and lower costs over time. We say more about these dynamic aspects of rivalry in optional Web Chapter 13.

LAST WORD

Internet Oligopolies

A Few Big Companies Dominate the Internet—and Act as Highly Competitive Oligopolists.

The Internet only became accessible to the average person in the mid-1990s. Over the past 10 years, it has evolved into a medium dominated by a few major firms. Chief among them are Google, Facebook, and Amazon. Other major players include Microsoft and Apple.

A key characteristic of each of these firms is that it holds a near-monopoly in a particular part of the tech business. Google dominates search. Facebook holds sway in social networking. Amazon runs the roost in online shopping. Microsoft holds a near-monopoly on PC operating systems and business-productivity software. And Apple became the world's most valuable company in 2012 by way of being the planet's most profitable manufacturer of computers, mobile phones, and tablets—all of which run on Apple's own operating software.

But instead of just trying to maintain dominance in its own sector, each of these Internet titans has used the profits generated by its own near-monopoly to try to steal business from one or more of the other titans. The result has been intense oligopolistic competition between a few well-funded rivals.

Consider search. Google's nearly 70 percent share of the search market creates massive amounts of advertising revenue for Google. In fact, Google's 2012 ad revenues of $20 billion exceeded the ad revenues received by all U.S. magazines and newspapers combined. So it may not be surprising that Microsoft created its Bing search engine to compete with Google. As of late 2012, Bing held 16 percent of the search market. Along with Yahoo, which held 12 percent, Bing maintains competitive pressure on Google, forcing ad rates lower.

Facebook is by far the largest social networking website, with more than 1 billion total users and 700 million regular users. But in 2011, Google succeeded in creating a large enough social network to challenge Facebook. Google did so by encouraging the users of its various free services—such as Gmail and YouTube—to join the Google+ social network. By late 2012, Google+ had 500 million total users and 235 million regular users—enough to compete credibly with Facebook.

Google+ was important for Google because Facebook had been encouraging advertisers to switch from using Google search ads to using Facebook banner ads that could be targeted at specific types of Facebook users (such as, "25–30 year old males with pets living in Pittsburgh"). Google can now counter by offering its own social network on which advertisers can place those sorts of targeted ads.

Google has also challenged Apple by releasing its very popular Android operating system for mobile devices to compete with the iOS operating system that Apple uses on both its iPhone cell phones and its iPad tablet computers. By doing so, Google reduced the threat that Apple could at some point in the future substantially reduce Google's search revenues by directing searches done on Apple devices to a proprietary search engine of Apple's own design.

Apple's dominance in smart phones and tablets has also been challenged by some of the other Internet titans. In addition to licensing the Android operating system to any manufacturer who wants to use it on their own cell phones or tablets, Google launched its own line of Nexus mobile devices to compete with Apple's iPhone and iPad. Also seeking to challenge Apple in mobile devices, Microsoft updated its Windows operating system to handle phones and tablets, launched its Surface line of tablets to compete with the iPad, and attempted to compete with the iPhone by marketing its own Windows Phone as well as by purchasing long-time cell-phone maker Nokia.

Microsoft's fundamental problem is that smartphone and tablet sales are rising rapidly while PC sales are falling quickly. So unless Microsoft can generate revenues from smartphones, tablets, or search, it will suffer an inexorable decline as the PC sector continues to shrink.

Amazon has also made forays outside of the online retail sector that it dominates. The foray best known to the general public is its Kindle line of tablet computers, which compete directly with the tablets made by Apple, Google, Microsoft, and the various manufactures that utilize Google's Android operating system. But behind the scenes, Amazon has also become a major competitor to Google and Microsoft in providing businesses with online "cloud computing" services that run on the massive servers that Amazon, Google, and Microsoft must maintain for their core businesses (like search). So in cloud computing, too, we see oligopoly competition resulting from Internet titans branching out of their own dominant sectors to compete with each other.

There's a simple reason for their aggressive competition. When a near-monopoly already dominates its own sector, its only chance for major profit growth is to invade a rival's sector.

SUMMARY

LO13.1 List the characteristics of monopolistic competition.

The distinguishing features of monopolistic competition are (*a*) there are enough firms in the industry to ensure that each firm has only limited control over price, mutual interdependence is absent, and collusion is nearly impossible; (*b*) products are characterized by real or perceived differences so that economic rivalry entails both price and nonprice competition; and (*c*) entry to the industry is relatively easy. Many aspects of retailing, and some manufacturing industries in which economies of scale are few, approximate monopolistic competition.

The four-firm concentration ratio measures the percentage of total industry output accounted for by the largest four firms. The Herfindahl index sums the squares of the percent market shares of all firms in the industry.

LO13.2 Explain why monopolistic competitors earn only a normal profit in the long run.

Monopolistically competitive firms may earn economic profits or incur losses in the short run. The easy entry and exit of firms results in only normal profits in the long run.

LO13.3 Explain why monopolistic competition delivers neither productive nor allocative efficiency.

The long-run equilibrium position of the monopolistically competitive producer is less efficient than that of the pure competitor. Under monopolistic competition, price exceeds marginal cost, indicating an underallocation of resources to the product, and price exceeds minimum average total cost, indicating that consumers do not get the product at the lowest price that cost conditions might allow.

LO13.4 Relate how the ability of monopolistic competition to deliver product differentiation helps to compensate for its failure to deliver economic efficiency.

Nonprice competition provides a way that monopolistically competitive firms can offset the long-run tendency for economic profit to fall to zero. Through product differentiation, product development, and advertising, a firm may strive to increase the demand for its product more than enough to cover the added cost of such nonprice competition. Consumers benefit from the wide diversity of product choice that monopolistic competition provides.

LO13.5 Describe the characteristics of oligopoly.

In practice, the monopolistic competitor seeks the specific combination of price, product, and advertising that will maximize profit.

Oligopolistic industries are characterized by the presence of few firms, each having a significant fraction of the market. Firms thus situated engage in strategic behavior and are mutually interdependent: The behavior of any one firm directly affects, and is affected by, the actions of rivals. Products may be either virtually uniform or significantly differentiated. Various barriers to entry, including economies of scale, underlie and maintain oligopoly.

High concentration ratios are an indication of oligopoly (monopoly) power. By giving more weight to larger firms, the Herfindahl index is designed to measure market dominance in an industry.

LO13.6 Discuss how game theory relates to oligopoly.

Game theory (*a*) shows the interdependence of oligopolists' pricing policies, (*b*) reveals the tendency of oligopolists to collude, and (*c*) explains the temptation of oligopolists to cheat on collusive arrangements.

LO13.7 Explain the three main models of oligopoly pricing and output: kinked-demand theory, collusive pricing, and price leadership.

Noncollusive oligopolists may face a kinked-demand curve. This curve and the accompanying marginal-revenue curve help explain the price rigidity that often characterizes oligopolies; they do not, however, explain how the actual prices of products were first established.

The uncertainties inherent in oligopoly promote collusion. Collusive oligopolists such as cartels maximize joint profits—that is, they behave like pure monopolists. Demand and cost differences, a "large" number of firms, cheating through secret price concessions, recessions, and the antitrust laws are all obstacles to collusive oligopoly.

Price leadership is an informal means of collusion whereby one firm, usually the largest or most efficient, initiates price changes and the other firms in the industry follow the leader.

LO13.8 Contrast the potential positive and negative effects of advertising.

Market shares in oligopolistic industries are usually determined on the basis of product development and advertising. Oligopolists emphasize nonprice competition because (*a*) advertising and product variations are less easy for rivals to match and (*b*) oligopolists frequently have ample resources to finance nonprice competition.

Advertising may affect prices, competition, and efficiency either positively or negatively. Positive: It can provide consumers with low-cost information about competing products, help introduce new competing products into concentrated industries, and generally reduce monopoly power and its attendant inefficiencies. Negative: It can promote monopoly power via persuasion and the creation of entry barriers. Moreover, it can be self-canceling when engaged in by rivals; then it boosts costs and creates inefficiency while accomplishing little else.

LO13.9 Discuss the efficiency of oligopoly from society's standpoint and whether it is more or less efficient than monopoly.

Neither productive nor allocative efficiency is realized in oligopolistic markets, but oligopoly may be superior to pure competition in promoting research and development and technological progress.

Table 10.1, page 221, provides a concise review of the characteristics of monopolistic competition and oligopoly as they compare to those of pure competition and pure monopoly.

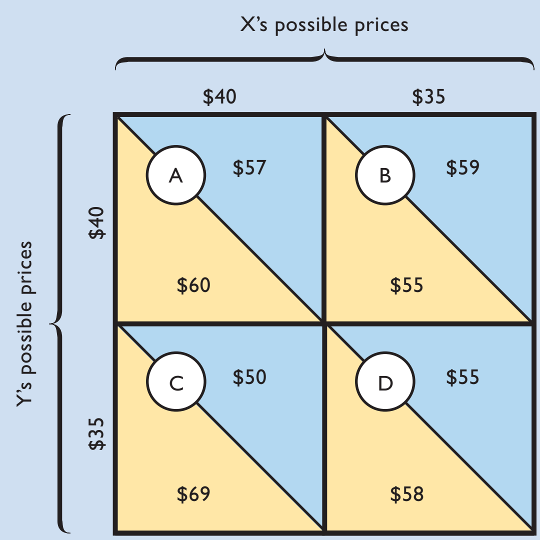
TERMS AND CONCEPTS

monopolistic competition
product differentiation
nonprice competition
four-firm concentration ratio
Herfindahl index
excess capacity
oligopoly
homogeneous oligopoly
differentiated oligopoly
strategic behavior
mutual interdependence
interindustry competition
import competition
game theory
collusion
kinked-demand curve
price war
cartel
price leadership

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. How does monopolistic competition differ from pure competition in its basic characteristics? From pure monopoly? Explain fully what product differentiation may involve. Explain how the entry of firms into its industry affects the demand curve facing a monopolistic competitor and how that, in turn, affects its economic profit. **LO13.1**
2. Compare the elasticity of a monopolistic competitor's demand with that of a pure competitor and a pure monopolist. Assuming identical long-run costs, compare graphically the prices and outputs that would result in the long run under pure competition and under monopolistic competition. Contrast the two market structures in terms of productive and allocative efficiency. Explain: "Monopolistically competitive industries are populated by too many firms, each of which produces too little." **LO13.2**
3. "Monopolistic competition is monopolistic up to the point at which consumers become willing to buy close-substitute products and competitive beyond that point." Explain. **LO13.2**
4. "Competition in quality and service may be just as effective as price competition in giving buyers more for their money." Do you agree? Why? Explain why monopolistically competitive firms frequently prefer nonprice competition to price competition. **LO13.2**
5. Critically evaluate and explain: **LO13.2**
   a. In monopolistically competitive industries, economic profits are competed away in the long run; hence, there is no valid reason to criticize the performance and efficiency of such industries.
   b. In the long run, monopolistic competition leads to a monopolistic price but not to monopolistic profits.
6. Why do oligopolies exist? List five or six oligopolists whose products you own or regularly purchase. What distinguishes oligopoly from monopolistic competition? **LO13.5**
7. Answer the following questions, which relate to measures of concentration: **LO13.5**
   a. What is the meaning of a four-firm concentration ratio of 60 percent? 90 percent? What are the shortcomings of concentration ratios as measures of monopoly power?
   b. Suppose that the five firms in industry A have annual sales of 30, 30, 20, 10, and 10 percent of total industry sales. For the five firms in industry B, the figures are 60, 25, 5, 5, and 5 percent. Calculate the Herfindahl index for each industry and compare their likely competitiveness.
8. Explain the general meaning of the profit payoff matrix below for oligopolists X and Y. All profit figures are in thousands. **LO13.6**
   a. Use the payoff matrix to explain the mutual interdependence that characterizes oligopolistic industries.
   b. Assuming no collusion between X and Y, what is the likely pricing outcome?
   c. In view of your answer to 8b, explain why price collusion is mutually profitable. Why might there be a temptation to cheat on the collusive agreement?



9. What assumptions about a rival's response to price changes underlie the kinked-demand curve for oligopolists? Why is there a gap in the oligopolists' marginal-revenue curve? How does the kinked-demand curve explain price rigidity in oligopoly? What are the shortcomings of the kinked-demand model? **LO13.7**
10. Why might price collusion occur in oligopolistic industries? Assess the economic desirability of collusive pricing. What are the main obstacles to collusion? Speculate as to why price leadership is legal in the United States, whereas price-fixing is not. **LO13.7**
11. Why is there so much advertising in monopolistic competition and oligopoly? How does such advertising help consumers and promote efficiency? Why might it be excessive at times? **LO13.8**
12. **ADVANCED ANALYSIS** Construct a game-theory matrix involving two firms and their decisions on high versus low advertising budgets and the effects of each on profits. Show a circumstance in which both firms select high advertising budgets even though both would be more profitable with low advertising budgets. Why won't they unilaterally cut their advertising budgets? **LO13.8**
13. **LAST WORD** Why have tech firms with near monopolies in their own sectors sought to compete with tech firms that have extremely strong, near-monopoly positions in other sectors?

REVIEW QUESTIONS

1. There are 10 firms in an industry, and each firm has a market share of 10 percent. The industry's Herfindahl index is: **LO13.1**
   a. 10.
   b. 100.
   c. 1,000.
   d. 10,000.
2. In the small town of Geneva, there are 5 firms that make watches. The firms' respective output levels are 30 watches per year, 20 watches per year, 20 watches per year, 20 watches per year, and 10 watches per year. The four-firm concentration ratio for the town's watch-making industry is: **LO13.1**
   a. 5.
   b. 70.
   c. 90.
   d. 100.
3. Which of the following best describes the efficiency of monopolistically competitive firms? **LO13.3**
   a. Allocatively efficient but productively inefficient.
   b. Allocatively inefficient but productively efficient.
   c. Both allocatively efficient and productively efficient.
   d. Neither allocatively efficient nor productively efficient.
4. Which of the following apply to oligopoly industries? Select one or more answers from the choices shown. **LO13.5**
   a. A few large producers.
   b. Many small producers.
   c. Strategic behavior.
   d. Price taking.
5. Faceblock, Gargle+, and MyMace are rival firms in an oligopoly industry. If kinked-demand theory applies to these three firms, Faceblock's demand curve will be: **LO13.7**
   a. More elastic above the current price than below it.
   b. Less elastic above the current price than below it.
   c. Of equal elasticity both above and below the current price.
   d. None of the above.
6. Consider an oligopoly industry whose firms have identical demand and cost conditions. If the firms decide to collude, then they will want to collectively produce the amount of output that would be produced by: **LO13.7**
   a. A monopolistic competitor.
   b. A pure competitor.
   c. A pure monopolist.
   d. None of the above.
7. In an oligopoly, each firm's share of the total market is typically determined by: **LO13.8**
   a. Scarcity and competition.
   b. Kinked demand curves and payoff matrices.
   c. Homogeneous products and import competition.
   d. Product development and advertising.
8. Some analysts consider oligopolies to be potentially less efficient than monopoly firms because at least monopoly firms tend to be regulated. Arguments in favor of a more benign view of oligopolies include: **LO13.9**
   a. Oligopolies are self-regulating.
   b. Oligopolies can be kept in line by foreign competition.
   c. Oligopolistic industries may promote technological progress.
   d. Oligopolies may engage in limit pricing to keep out potential entrants.

PROBLEMS

1. Suppose that a small town has seven burger shops whose respective shares of the local hamburger market are (as percentages of all hamburgers sold): 23%, 22%, 18%, 12%, 11%, 8%, and 6%. What is the four-firm concentration ratio of the hamburger industry in this town? What is the Herfindahl index for the hamburger industry in this town?

If the top three sellers combined to form a single firm, what would happen to the four-firm concentration ratio and to the Herfindahl index? **LO13.5**

2. Suppose that the most popular car dealer in your area sells 10 percent of all vehicles. If all other car dealers sell either the same number of vehicles or fewer, what is the largest value that the Herfindahl index could possibly take for car dealers in your area? In that same situation, what would the four-firm concentration ratio be? **LO13.5**
3. Suppose that an monopolistically competitive restaurant is currently serving 230 meals per day (the output where MR = MC). At that output level, ATC per meal is $10 and consumers are willing to pay $12 per meal. What is the size of this firm's profit or loss? Will there be entry or exit? Will this restaurant's demand curve shift left or right? In long-run equilibrium, suppose that this restaurant charges $11 per meal for 180 meals and that the marginal cost of the 180th meal is $8. What is the size of the firm's profit? Suppose that the allocatively efficient output level in long-run equilibrium is 200 meals. Is the deadweight loss for this firm greater than or less than $60? **LO13.5**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

Additional Game Theory Applications

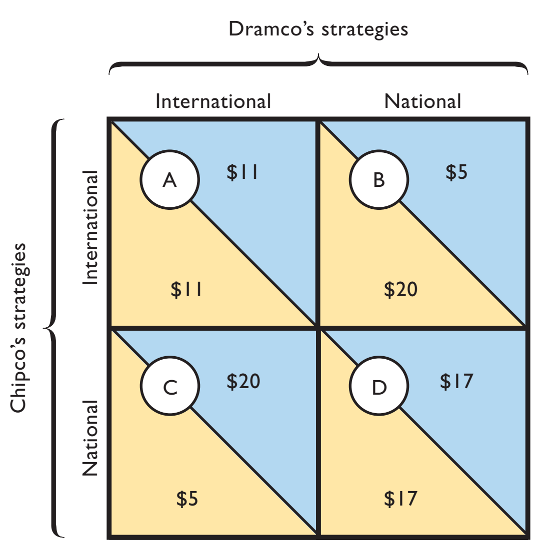
LO13.10 Utilize additional game-theory terminology and applications.

We have seen that game theory is helpful in explaining mutual interdependence and strategic behavior by oligopolists. This appendix provides additional oligopoly-based applications of game theory.

A One-Time Game: Strategies and Equilibrium

Consider Figure 1, which lists strategies and outcomes for two fictitious producers of the computer memory chips referred to as DRAMs (Dynamic Random Access Memory circuits). Chipco is the single producer of these chips in the United States and Dramco is the only producer in China. Each firm has two alternative strategies: an international strategy, in which it competes directly against the other firm in both countries; and a national strategy, in which it sells only in its home country.

FIGURE 1 A One-Time Game In this single-period, positive-sum game, Chipco's international strategy is its dominant strategy—the alternative that is superior to any other strategy regardless of whatever Dramco does. Similarly, Dramco's international strategy is also its dominant strategy. With both firms choosing international strategies, the outcome of the game is Cell A, where each firm receives an $11 million profit. Cell A is a Nash equilibrium because neither firm will independently want to move away from it given the other firm's strategy.



The game and payoff matrix shown in Figure 1 is a **one-time game** because the firms select their optimal strategies in a single time period without regard to possible interactions in subsequent time periods. The game is also a **simultaneous game** because the firms choose their strategies at the same time; and a **positive-sum game,** a game in which the sum of the two firms' outcomes (here, profits) is positive. In contrast, the net gain in a **zero-sum game** is zero because one firm's gain must equal the other firm's loss, and the net gain in a **negative-sum game** is negative. In some positive-sum games, both firms may have positive outcomes. That is the case in Figure 1.

To determine optimal strategies, Chipco looks across the two rows in the payoff matrix (yellow portion of cells in millions of dollars) and Dramco looks down the two columns (blue portion of cells). These payoffs indicate that both firms have a **dominant strategy**—an option that is better than any alternative option *regardless of what the other firm does.* To see this, notice that Chipco's international strategy will give it a higher profit than its national strategy—regardless of whether Dramco chooses to utilize an international or a national strategy. An international strategy will produce an $11 million profit for Chipco (yellow portion of cell A) if Dramco also uses an international strategy while a national strategy will result in a $20 million profit for Chipco (yellow portion of cell B) if Dramco uses a national strategy. Chipco's possible $11 million and $20 million outcomes are clearly better than the $5 million (cell C) and $17 million (cell D) outcomes it could receive if it chose to pursue a national strategy. Chipco's international strategy is, consequently, its dominant strategy. Using similar logic, Dramco also concludes that its international strategy is its dominant strategy.

In this particular case, the outcome (cell A) of the two dominant strategies is the game's **Nash equilibrium**—an outcome from which neither rival wants to deviate.[1] At the Nash equilibrium, both rivals see their current strategy as optimal *given the other firm's strategic choice.* The

[1]The Nash equilibrium is named for its discoverer, John F. Nash. Nash's life and Nobel Prize are the subject of the motion picture *A Beautiful Mind*, directed by Ron Howard and starring Russell Crowe.

Nash equilibrium is the only outcome in the payoff matrix in Figure 1 that, once achieved, is stable and therefore will persist.[2]

Credible and Empty Threats

In looking for optimal strategies, Chipco and Dramco both note that they could increase their profit from $11 million to $17 million if they could agree to jointly pursue national strategies (cell D) instead of independently pursuing international strategies (cell A). Presumably the national strategies would leave the firms as pure monopolists in their domestic economies, with each able to set higher prices and obtain greater profits as a result. But if this territorial agreement were put in place, both firms would have an incentive to cheat on the agreement by secretly selling DRAMs in the other's country. That would temporarily move the game to either cell B or cell C. Once discovered, however, such cheating would undermine the territorial agreement and return the game to the Nash equilibrium (cell A).

Now let's add a new twist—a credible threat—to the game shown in Figure 1. A **credible threat** is a statement of coercion (a threat!) that is believable by the other firm. Suppose that Chipco is the lower-cost producer of DRAMs because of its superior technology. Also, suppose that Chipco approaches Dramco saying that Chipco intends to use its national strategy and expects Dramco to do the same. If Dramco decides against the national strategy or agrees to the strategy and then later cheats on the agreement, Chipco will immediately drop its price to an ultra-low level equal to its average total cost (ATC). Both firms know that Chipco's ATC price is below Dramco's ATC. Although Chipco will see its economic profit fall to zero, Dramco will suffer an economic loss and possibly go out of business.

If Chipco's threat is credible, the two firms represented in Figure 1 will abandon the Nash equilibrium (cell A) to deploy their national strategies and achieve highly profitable cell D. In game theory, credible threats such as this can help establish and maintain collusive agreements. A strong "enforcer" can help prevent cheating and maintain the group discipline needed for cartels, price-fixing conspiracies, and territorial understandings to successfully generate high profits.

But credible threats are difficult to achieve in the actual economy. For example, Dramco might rightly wonder why Chipco had not previously driven it out of business through an ultra-low price strategy. Is Chipco fearful of the U.S. antitrust authorities?

[2]Nash equilibriums can exist even in games that lack dominant strategies.

If Dramco does not wish to participate in the proposed scheme, it might counter Chipco's threat with its own: Forget that you ever talked to us and we will not take this illegal "offer" to the U.S. Justice Department. Dramco can make this threat because strict laws are in place against attempts to restrain trade through price-fixing and territorial agreements.

So Dramco may view Chipco's threat as simply an **empty threat**—a statement of coercion that is not believable by the threatened firm. If so, the Nash equilibrium will prevail, with both firms pursuing an international strategy.

Repeated Games and Reciprocity Strategies

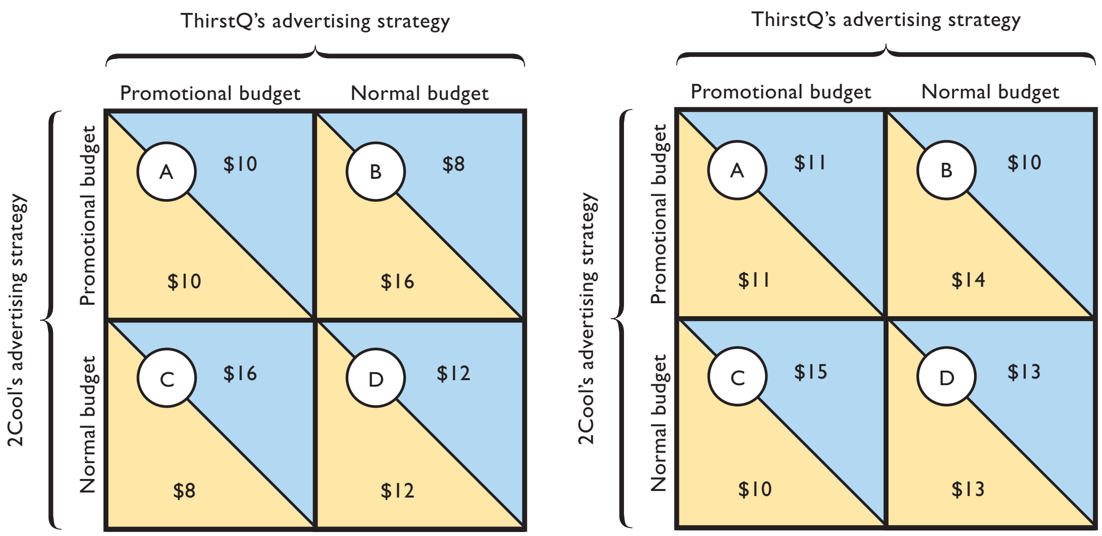
The Chipco-Dramco game was a one-time game, but many strategic situations are repeated by the same oligopolists over and over again. For example, Coca Cola and Pepsi are mutually interdependent on pricing, advertising, and product development year after year, decade after decade. The same is true for Boeing and Airbus, Walmart and Target, Toyota and General Motors, Budweiser and Miller, Nike and Adidas, and numerous other dominant pairs.

In a **repeated game**—a game that recurs more than once—the optimal strategy may be to cooperate and restrain oneself from competing as hard as possible so long as the other firm reciprocates by also not competing as hard as possible.[3] To see how this works, consider two hypothetical producers of soft drinks: 2Cool and ThirstQ. If ThirstQ competes hard with 2Cool in today's situation in which 2Cool would like ThirstQ to take things easy, 2Cool will most likely retaliate against ThirstQ in any subsequent situation where the circumstances are reversed. In contrast, if ThirstQ cooperates with 2Cool in game 1, ThirstQ can expect 2Cool to reciprocate in game 2 of their repeated interaction. Both firms know full well the negative long-run consequences of ever refusing to cooperate. So the cooperation continues, not only in game 2, but in games 3, 4, 5, and beyond.

Figure 2 shows two side-by-side payoff matrixes for the two games. In Figure 2a, 2Cool and ThirstQ face a situation in which 2Cool is introducing a new cola called Cool Cola and has two advertising options: a high promotional budget to introduce the new product and a normal

[3]We are assuming either an infinitely repeated game or a game of unknown time-horizon. Games with a known ending date undermine reciprocity strategies.

FIGURE 2 A Repeated Game with Reciprocity (a) in the payoff matrix to the left, 2Cool introduces its new Cool Cola with a large promotional advertising budget, but its rival ThirstQ maintains its normal advertising budget even though it could counter 2Cool with a large advertising budget of its own and drive the outcome from Cell B to Cell A. ThirstQ forgoes this $2 million of extra profit because it knows that it will soon be introducing its own new product (Quench It). (b) In the payoff matrix to the right, ThirstQ introduces Quench It with a large promotional advertising budget. Cool2 reciprocates ThirstQ's earlier accommodation by not matching ThirstQ's promotional advertising budget and instead allowing the outcome of the repeated game to be Cell C. The profit of both 2Cool and ThirstQ therefore is larger over the two periods than if each firm had aggressively countered each other's single-period strategy.



(a) 2Cool introduces Cool Cola

(b) ThirstQ introduces Quench It

advertising budget. ThirstQ has the same two options: a high promotional budget to try to counter 2Cool's product introduction and a normal advertising budget.

The analysis is now familiar to you. The dominant strategies for both firms in game 1 (Figure 2a) are their large promotional advertising budgets and the Nash equilibrium is cell A. Both firms could do better at cell D if each agreed to use normal advertising budgets. But 2Cool could do better still. It could achieve the $16 million of profit in cell B, but only if ThirstQ holds its advertising budget to its normal level during the introduction of Cool Cola.

ThirstQ might voluntarily do just that! It knows that game 2 (Figure 2b) is forthcoming in which it will be introducing its new product, Quench It. By leaving its advertising budget at its normal level during 2Cool's introduction of Cool Cola, and thereby sacrificing profit of $2 million (= $10 million in cell A − $8 million in cell B), ThirstQ can expect 2Cool to reciprocate in the subsequent game in which ThirstQ introduces Quench It.

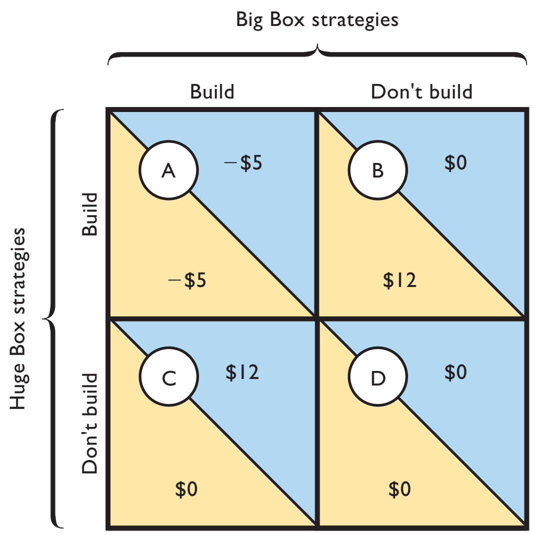
Without formally colluding—and risking antitrust penalties—game 1 ends at cell B and repeated game 2 ends at cell C. With reciprocity, 2Cool's total profit of $26 million (= $16 million in game 1 + $10 million in game 2) exceeds the $21 million (= $10 million + $11 million) it would have earned without the reciprocity. ThirstQ similarly benefits. To check your understanding, confirm this fact using the numbers in the two matrixes.

First-Mover Advantages and Preemption of Entry

The games we have highlighted thus far have been games in which the two firms simultaneously select their optimal strategies. But in some actual economic circumstances, firms apply strategies sequentially: One firm moves first and commits to a strategy to which a rival firm must subsequently respond. In such a **sequential game,** the final outcome may depend critically upon which firm moves first since the first mover may have the opportunity to establish a Nash equilibrium that works in its favor.

Consider Figure 3, which identifies a game in which two large retailers—let's call them Big Box and Huge Box—are each considering building a large retail store in a small rural city. As indicated in the figure, each firm has two strategies: build or don't build. The payoff matrix reflects the fact that the city is not large enough to support two big box retailers profitably. If both retailers simultaneously build,

FIGURE 3 A First-Mover Advantage and the Preemption of Entry
In this game in which strategies are pursued sequentially, the firm that moves first can take advantage of the particular situation represented in which only a single firm can exist profitably in some geographical market. Here, we suppose that Big Box moves first with its "Build" strategy to achieve the $12 million profit outcome in Cell C. Huge Box then will find that it will lose money if it also builds because that will result in a $5 million loss, as shown in Cell A.



the outcome will be cell A and each firm will lose $5 million. If neither firm builds, the outcome will be cell D with both firms securing zero profit. If only Big Box builds, the outcome will be cell C and Big Box will profit handsomely at $12 million. If Huge Box builds, but Big Box stays out, the outcome will be cell B and Huge Box will secure the $12 million profit. Either cell B or cell C is the possible Nash equilibrium. At either cell, both firms will have selected their best option in view of the strategy taken by the other firm.

The payoff matrix in Figure 3 clearly reveals that whoever builds first will preempt the other retailer from entering the market. An extremely large **first-mover advantage** exists in this particular game. Suppose that a well-thought-out strategy and adequate financing leave Big Box better prepared than Huge Box to move quickly to build a large retail store in this city. By exploiting its first-mover advantage, Big Box drives the outcome to Cell C and preempts Huge Box's entry into this market.

Many firms in the actual economy have used variations of this first-mover strategy to a greater or lesser extent to preempt major rivals, or at least greatly slow their entry. Examples are Walmart, Home Depot, Costco, Walgreens, Starbucks, and many more. The strategy, however, is highly risky because it requires the commitment of huge amounts of investment funds to saturate the market and preclude entry by other firms. Also, to be the first-mover in places that are being transformed from rural land into urban areas, firms may need to build their stores many months prior to the time when the area in question becomes developed enough to provide the store with significant business. That may mean losses until the market grows sufficiently for profitability. Some firms such as Walmart have become huge, profitable international enterprises by using a first-mover strategy. Other firms, such as Krispy Kreme Donuts, have lost millions of dollars because their extremely rapid expansion turned out to be unprofitable in many of their outlets because the expected customers never materialized.

APPENDIX SUMMARY

LO13.10 Utilize additional game-theory terminology and applications.

Positive-sum games are games in which the payoffs to the firms sum to a positive number; zero-sum games are games in which the payoffs sum to zero; and negative-sum games are games in which the payoffs sum to less than zero. Positive-sum games allow for "win-win" opportunities, whereas zero-sum games always feature "I win–you lose" outcomes. Games can be either one-time games or repeated games. Decisions in games may be made either simultaneously or sequentially.

When two firms are playing a strategic game, a firm is said to have a dominant strategy if there is an option that leads to better outcomes than all other options regardless of what the other firm does. Not all games have dominant strategies. The Nash equilibrium is an outcome from which neither firm wants to deviate because both firms see their current strategy as optimal given the other firm's chosen strategy. The Nash equilibrium is stable and persistent. Attempts by the firms to rig games to achieve some other outcome are difficult to accomplish and maintain, although credible threats can sometimes work. In contrast, empty threats accomplish nothing and leave the outcome at the Nash equilibrium.

Reciprocity can improve outcomes for firms participating in repeated games. In such games, one firm avoids taking advantage of the other firm because it knows that the other firm can take advantage of it in subsequent games. This reciprocity increases firm profits relative to what they would have been without reciprocity.

Two possible Nash equilibriums can exist in sequential games with first-mover advantages. Which one occurs depends on which firm moves first since that firm can preempt the other firm, making it unprofitable for the other firm to match the move. Several real-world firms, including Walmart, have successfully used first-mover advantages to saturate markets and preempt entry by rivals.

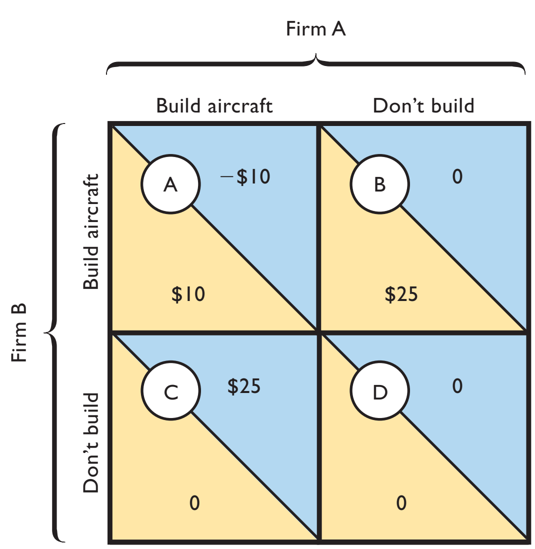
APPENDIX TERMS AND CONCEPTS

one-time game
simultaneous game
positive-sum game
zero-sum game
negative-sum game
dominant strategy
Nash equilibrium
credible threat
empty threat
repeated game
sequential game
first-mover advantage

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Is the game shown by Figure 13.3 in the chapter (not this appendix) a zero-sum game or is it a positive-sum game? How can you tell? Are there dominant strategies in this game? If so, what are they? What cell represents a Nash equilibrium and why? Explain why it is so difficult for Uptown and RareAir to achieve and maintain a more favorable cell than the Nash equilibrium in this single-period pricing game. **LO13.10**
2. Refer to the payoff matrix in discussion question 8 at the end of this chapter. First, assume this is a one-time game. Explain how the $60/$57 outcome might be achieved through a credible threat. Next, assume this is a repeated game (rather than a one-time game) and that the interaction between the two firms occurs indefinitely. Why might collusion with a credible threat not be necessary to achieve the $60/$57 outcome? **LO13.10**
3. Refer to the payoff matrix below. **LO13.10**



Assuming this is a sequential game with no collusion, what is the outcome if Firm A moves first to build a new type of commercial aircraft? Explain why first-mover strategies in the real world are only as good as the profit projections on which they are based. How could a supposed "win" from moving first turn out to be a big loss, whereas the "loss" of being preempted turn out to be a blessing in disguise?

4. **ADVANCED ANALYSIS** Suppose you are playing a game in which you and one other person each picks a number between 1 and 100, with the person closest to some randomly selected number between 1 and 100 winning the jackpot. (Ask your instructor to fund the jackpot.) Your opponent picks first. What number do you expect her to choose? Why? What number would you then pick? Why are the two numbers so close? How might this example relate to why Home Depot and Lowes, Walgreens and Rite-Aid, McDonald's and Burger King, and other major pairs of rivals locate so close to each other in many well-defined geographical markets that are large enough for both firms to be profitable? **LO13.10**

APPENDIX REVIEW QUESTIONS

1. Collusive agreements can be established and maintained by: **LO13.10**
   - **a.** Credible threats.
   - **b.** One-time games.
   - **c.** Empty threats.
   - **d.** First-mover advantage.
2. True or false. Potential rivals may be more likely to collude if they view themselves as playing a repeated game rather than a one-time game. **LO13.10**
3. Property developers who build shopping malls like to have them "anchored" with the outlets of one or more famous national retail chains, like Target or Nordstrom. Having such "anchors" is obviously good for the mall developers because anchor stores bring a lot of foot traffic that can help generate sales for smaller stores that lack well-known national brands. But what's in it for the national retail chains? Why become an anchor? Chose the best answer from the following list. **LO13.10**
   - **a.** The anchor stores want to make a credible threat against the developer.
   - **b.** The anchor stores may feel there is a first-mover advantage to becoming one of only a few anchor stores at a new mall.
   - **c.** The property developers are making empty threats to smaller stores.
   - **d.** The smaller stores face a negative-sum game.

APPENDIX PROBLEMS

1. Consider a "punishment" variation of the two-firm oligopoly situation shown in Figure 13.3 in the chapter (not in this appendix). Suppose that if one firm sets a low price while the other sets a high price, then the firm setting the high price can fine the firm setting the low price. Suppose that whenever a fine is imposed, X dollars is taken from the low-price firm and given to the high-price firm. What is the smallest amount that the fine X can be such that both firms will want to always set the high price? **LO13.10**
2. Consider whether the promises and threats made toward each other by duopolists and oligopolists are always credible (believable). Look back at Figure 13.3 in the chapter (not in this appendix). Imagine that the two firms will play this game twice in sequence and that each firm claims the following policy. Each says that if both it and the other firm choose the high price in the first game, then it will also choose the high price in the second game (as a reward to the other firm for cooperating in the first game). **LO13.10**
   - **a.** As a first step toward thinking about whether this policy is credible, consider the situation facing both firms in the second game. If each firm bases its decision on what to do in the second game entirely on the payouts facing the firms in the second game, which strategy will each firm choose in the second game?
   - **b.** Now move backward in time one step. Imagine that it is the start of the first game and each firm must decide what to do during the first game. Given your answer to 2a, is the publicly stated policy credible? (Hint: No matter what happens in the first game, what will both firms do in the second game?)
   - **c.** Given your answers to 2a and 2b, what strategy will each firm choose in the first game?

WEB CHAPTER 13

Technology, R&D, and Efficiency

www.mcconnell20e.com

Learning Objectives

LO13W.1 Differentiate between an invention, an innovation, and technological diffusion.

LO13W.2 Explain how entrepreneurs and other innovators further technological advance.

LO13W.3 Summarize how a firm determines its optimal amount of research and development (R&D).

LO13W.4 Discuss how technological change can increase profits by raising revenues or lowering costs.

LO13W.5 Relate why firms can benefit from their innovation even though rivals have an incentive to imitate it.

LO13W.6 Discuss the role of market structure in promoting technological advance.

LO13W.7 Show how technological advance enhances productive efficiency and allocative efficiency.

WEB CHAPTER

Web Chapter 13 is a bonus chapter found at the book's Web site, **www.mcconnell20e.com**. It extends the analysis of Part 4, "Microeconomics of Product Markets," by examining such topics as invention, innovation, R&D decision making, and creative destruction. Your instructor may (or may not) assign all or part of this chapter.

PART **FIVE**

MICROECONOMICS OF RESOURCE MARKETS AND GOVERNMENT

CHAPTER **14**

The Demand for Resources

Learning Objectives

LO14.1 **Explain the significance of resource pricing.**

LO14.2 **Convey how the marginal revenue productivity of a resource relates to a firm's demand for that resource.**

LO14.3 **List the factors that increase or decrease resource demand.**

LO14.4 **Discuss the determinants of elasticity of resource demand.**

LO14.5 **Determine how a competitive firm selects its optimal combination of resources.**

LO14.6 **Explain the marginal productivity theory of income distribution.**

When you finish your education, you probably will look for a new job. Employers have a demand for educated, productive workers like you. To learn more about the demand for labor and other resources, we now turn from the pricing and production of *goods and services* to the pricing and employment of *resources*. Although firms come in various sizes and operate under different market conditions, each has a demand for productive resources. Firms obtain needed resources from households—the direct or indirect owners of land, labor, capital, and entrepreneurial resources. We shift our attention from the bottom loop of the circular flow model (p. 43), where businesses supply products that households demand, to the top loop, where businesses demand resources that households supply.

This chapter looks at the *demand* for economic resources. Although the discussion is couched in terms of labor, the principles developed also apply to land, capital, and entrepreneurial ability. In Chapter 15 we will combine resource (labor) demand with labor *supply* to analyze wage rates. In Chapter 16 we will use resource demand and resource supply to examine the prices of, and returns to, other productive resources. Issues relating to the use of natural resources are the subject of Chapter 17.

Significance of Resource Pricing

LO14.1 Explain the significance of resource pricing.

Studying resource pricing is important for several reasons:

- ***Money-income determination*** Resource prices are a major factor in determining the income of households. The expenditures that firms make in acquiring economic resources flow as wage, rent, interest, and profit incomes to the households that supply those resources.
- ***Cost minimization*** To the firm, resource prices are costs. And to obtain the greatest profit, the firm must produce the profit-maximizing output with the most efficient (least costly) combination of resources. Resource prices play the main role in determining the quantities of land, labor, capital, and entrepreneurial ability that will be combined in producing each good or service (see Table 2.1, p. 39).
- ***Resource allocation*** Just as product prices allocate finished goods and services to consumers, resource prices allocate resources among industries and firms. In a dynamic economy, where technology and product demand often change, the efficient allocation of resources over time calls for the continuing shift of resources from one use to another. Resource pricing is a major factor in producing those shifts.
- ***Policy issues*** Many policy issues surround the resource market. Examples: To what extent should government redistribute income through taxes and transfers? Should government do anything to discourage "excess" pay to corporate executives? Should it increase the legal minimum wage? Is the provision of subsidies to farmers efficient? Should government encourage or restrict labor unions? The facts and debates relating to these policy questions are grounded on resource pricing.

Marginal Productivity Theory of Resource Demand

LO14.2 Convey how the marginal revenue productivity of a resource relates to a firm's demand for that resource.

In discussing resource demand, we will first assume that a firm sells its output in a purely competitive product market and hires a certain resource in a purely competitive resource market. This assumption keeps things simple and is consistent with the model of a competitive labor market that we will develop in Chapter 15. In a competitive *product market*, the firm is a "price taker" and can dispose of as little or as much output as it chooses at the market price. The firm is selling such a negligible fraction of total output that its output decisions exert no influence on product price. Similarly, the firm also is a "price taker" (or "wage taker") in the competitive *resource market*. It purchases such a negligible fraction of the total supply of the resource that its buying (or hiring) decisions do not influence the resource price.

Resource Demand as a Derived Demand

Resource demand is the starting point for any discussion of resource prices. Resource demand is a schedule or a curve showing the amounts of a resource that buyers are willing and able to purchase at various prices over some period of time. Crucially, resource demand is a **derived demand,** meaning that the demand for a resource is derived from the demand for the products that the resource helps to produce. This is true because resources usually do not directly satisfy customer wants but do so indirectly through their use in producing goods and services. Almost nobody wants to consume an acre of land, a John Deere tractor, or the labor services of a farmer, but millions of households do want to consume the food and fiber products that these resources help produce. Similarly, the demand for airplanes generates a demand for assemblers, and the demands for such services as income-tax preparation, haircuts, and child care create derived demands for accountants, barbers, and child care workers.

Marginal Revenue Product

Because resource demand is derived from product demand, the strength of the demand for any resource will depend on:

- The productivity of the resource in helping to create a good or service.
- The market value or price of the good or service it helps produce.

Other things equal, a resource that is highly productive in turning out a highly valued commodity will be in great demand. On the other hand, a relatively unproductive resource that is capable of producing only a minimally valued commodity will be in little demand. And no demand whatsoever will exist for a resource that is phenomenally efficient in producing something that no one wants to buy.

Productivity Table 14.1 shows the roles of resource productivity and product price in determining resource demand. Here we assume that a firm adds a single variable resource, labor, to its fixed plant. Columns 1 and 2 give the number of units of the resource applied to production and the resulting total product (output). Column 3 provides the **marginal product (MP),** or additional output, resulting

TABLE 14.1 The Demand for Labor: Pure Competition in the Sale of the Product

(1) Units of Resource	(2) Total Product (Output)	(3) Marginal Product (MP)	(4) Product Price	(5) Total Revenue, (2) × (4)	(6) Marginal Revenue Product (MRP)
0	0		$2	$ 0	
		7			$14
1	7		2	14	
		6			12
2	13		2	26	
		5			10
3	18		2	36	
		4			8
4	22		2	44	
		3			6
5	25		2	50	
		2			4
6	27		2	54	
		1			2
7	28		2	56	

from using each additional unit of labor. Columns 1 through 3 remind us that the law of diminishing returns applies here, causing the marginal product of labor to fall beyond some point. For simplicity, we assume that these diminishing marginal returns—these declines in marginal product—begin with the first worker hired.

Product Price But the derived demand for a resource depends also on the price of the product it produces. Column 4 in Table 14.1 adds this price information. Product price is constant, in this case at $2, because the product market is competitive. The firm is a price taker and can sell units of output only at this market price.

Multiplying column 2 by column 4 provides the total-revenue data of column 5. These are the amounts of revenue the firm realizes from the various levels of resource usage. From these total-revenue data we can compute **marginal revenue product (MRP)**—the change in total revenue resulting from the use of each additional unit of a resource (labor, in this case). In equation form,

$$\text{Marginal revenue product} = \frac{\text{change in total revenue}}{\text{unit change in resource quantity}}$$

The MRPs are listed in column 6 in Table 14.1.

Rule for Employing Resources: MRP = MRC

The MRP schedule, shown as columns 1 and 6, is the firm's demand schedule for labor. To understand why, you must first know the rule that guides a profit-seeking firm in hiring any resource: To maximize profit, a firm should hire additional units of a specific resource as long as each successive unit adds more to the firm's total revenue than it adds to the firm's total cost.

Economists use special terms to designate what each additional unit of labor or other variable resource adds to total cost and what it adds to total revenue. We have seen that MRP measures how much each successive unit of a resource adds to total revenue. The amount that each additional unit of a resource adds to the firm's total (resource) cost is called its **marginal resource cost (MRC).** In equation form,

$$\text{Marginal resource cost} = \frac{\text{change in total (resource) cost}}{\text{unit change in resource quantity}}$$

So we can restate our rule for hiring resources as follows: It will be profitable for a firm to hire additional units of a resource up to the point at which that resource's MRP is equal to its MRC. For example, as the rule applies to labor, if the number of workers a firm is currently hiring is such that the MRP of the last worker exceeds his or her MRC, the firm can profit by hiring more workers. But if the number being hired is such that the MRC of the last worker exceeds his or her MRP, the firm is hiring workers who are not "paying their way" and it can increase its profit by discharging some workers. You may have recognized that this **MRP = MRC rule** is similar to the MR = MC profit-maximizing rule employed throughout our discussion of price and output determination. The rationale of the two rules is the same, but the point of reference is now *inputs* of a resource, not *outputs* of a product.

MRP as Resource Demand Schedule

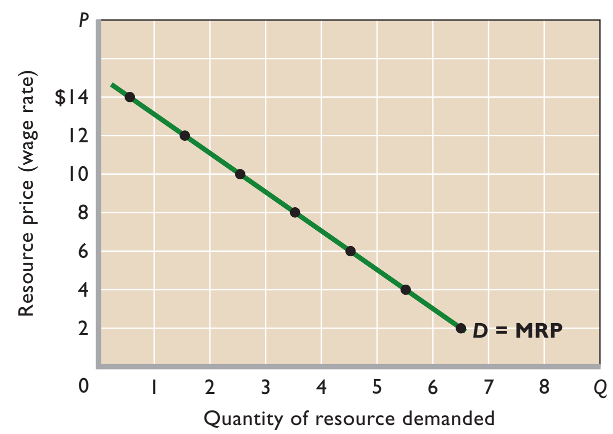
Let's continue with our focus on labor, knowing that the analysis also applies to other resources. In a purely competitive labor market, market supply and market demand establish the wage rate. Because each firm hires such a small fraction of market supply, it cannot influence the market wage rate; it is a wage taker, not a wage maker. This means that for each additional unit of labor hired, each firm's total resource cost increases by exactly the amount of the constant market wage rate. More specifically, the MRC of labor exactly equals the market wage rate. Thus, resource "price" (the market wage rate) and resource "cost" (marginal resource cost) are equal for a firm that hires a resource in a competitive labor market. As a result, the MRP = MRC rule tells us that, in pure competition, the firm will hire workers up to the point at which the market *wage rate* (its MRC) is equal to its MRP.

In terms of the data in columns 1 and 6 of Table 14.1, if the market wage rate is, say, $13.95, the firm will hire only one worker. This is so because only the hiring of the first worker results in an increase in profits. To see this, note that for the first worker MRP (= $14) exceeds MRC (= $13.95). Thus, hiring the first worker is profitable. For each successive worker, however, MRC (= $13.95) exceeds MRP (= $12 or less), indicating that it will not be profitable to hire any of those workers. If the wage rate is $11.95, by the same reasoning we discover that it will pay the firm to hire both the first and second workers. Similarly, if the wage rate is $9.95, three workers will be hired. If it is $7.95, four. If it is $5.95, five. And so forth. So here is the key generalization: The MRP schedule constitutes the firm's demand for labor because each point on this schedule (or curve) indicates the number of workers the firm would hire at each possible wage rate.

In Figure 14.1, we show the *D* = MRP curve based on the data in Table 14.1.[1] The competitive firm's resource demand curve identifies an inverse relationship between the wage rate and the quantity of labor demanded, other things equal. The curve slopes downward because of diminishing marginal returns.

[1]Note that we plot the points in Figure 14.1 halfway between succeeding numbers of resource units because MRP is associated with the addition of 1 more unit. Thus in Figure 14.1, for example, we plot the MRP of the second unit ($12) not at 1 or 2 but at $1\frac{1}{2}$. This "smoothing" enables us to sketch a continuously downsloping curve rather than one that moves downward in discrete steps (like a staircase) as each new unit of labor is hired.

FIGURE 14.1 The purely competitive seller's demand for a resource. The MRP curve is the resource demand curve; each of its points relates a particular resource price (= MRP when profit is maximized) with a corresponding quantity of the resource demanded. Under pure competition, product price is constant; therefore, the downward slope of the *D* = MRP curve is due solely to the decline in the resource's marginal product (law of diminishing marginal returns).



Resource Demand under Imperfect Product Market Competition

Resource demand (here, labor demand) is more complex when the firm is selling its product in an imperfectly competitive market, one in which the firm is a price maker. That is because imperfect competitors (pure monopolists, oligopolists, and monopolistic competitors) face downsloping product demand curves. As a result, whenever an imperfect competitor's product demand curve is fixed in place, the only way to increase sales is by setting a lower price (and thereby moving down along the fixed demand curve).

The productivity data in Table 14.1 are retained in columns 1 to 3 in Table 14.2. But here in Table 14.2 we show in column 4 that product price must be lowered to sell the

TABLE 14.2 The Demand for Labor: Imperfect Competition in the Sale of the Product

(1) Units of Resource	(2) Total Product (Output)	(3) Marginal Product (MP)	(4) Product Price	(5) Total Revenue, (2) × (4)	(6) Marginal Revenue Product (MRP)
0	0		$2.80	$ 0	
		7			$ 18.20
1	7		2.60	18.20	
		6			13.00
2	13		2.40	31.20	
		5			8.40
3	18		2.20	39.60	
		4			4.40
4	22		2.00	44.00	
		3			2.25
5	25		1.85	46.25	
		2			1.00
6	27		1.75	47.25	
		1			−1.05
7	28		1.65	46.20	

marginal product of each successive worker. The MRP of the purely competitive seller of Table 14.1 falls for only one reason: Marginal product diminishes. But the MRP of the imperfectly competitive seller of Table 14.2 falls for two reasons: Marginal product diminishes *and* product price falls as output increases.

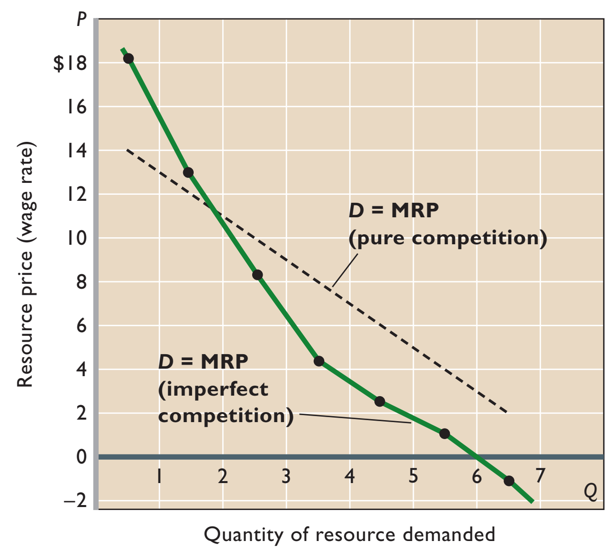
We emphasize that the lower price accompanying each increase in output (total product) applies not only to the marginal product of each successive worker but also to all prior output units that otherwise could have been sold at a higher price. Observe that the marginal product of the second worker is 6 units of output. These 6 units can be sold for $2.40 each, or, as a group, for $14.40. But $14.40 is not the MRP of the second worker. To sell these 6 units, the firm must take a 20-cent price cut on the 7 units produced by the first worker—units that otherwise could have been sold for $2.60 each. Thus, the MRP of the second worker is only $13 [= $14.40 − (7 × 20 cents)], as shown.

Similarly, the third worker adds 5 units to total product, and these units are worth $2.20 each, or $11 total. But to sell these 5 units, the firm must take a 20-cent price cut on the 13 units produced by the first two workers. So the third worker's MRP is only $8.40 [= $11 − (13 × 20 cents)]. The numbers in column 6 reflect such calculations.

In Figure 14.2 we graph the MRP data from Table 14.2 and label it "*D* = MRP (imperfect competition)." The broken-line resource demand curve, in contrast, is that of the purely competitive seller represented in Figure 14.1. A comparison of the two curves demonstrates that, other things equal, the resource demand curve of an imperfectly competitive seller is less elastic than that of a purely competitive seller. Consider the effects of an identical percentage decline in the wage rate (resource price) from $11 to $6 in Figure 14.2. Comparison of the two curves reveals that the imperfectly competitive seller (solid curve) does not expand the quantity of labor it employs by as large a percentage as does the purely competitive seller (broken curve).

It is not surprising that the imperfectly competitive producer is less responsive to resource price cuts than the purely competitive producer. When resource prices fall, MC per unit declines for both imperfectly competitive firms as well as purely competitive firms. Because both types of firms maximize profits by producing where MR = MC, the decline in MC will cause both types of firms to produce more. But the effect will be muted for imperfectly competitive firms because their downsloping demand curves cause them to also face downsloping MR curves—so that for each additional unit sold, MR declines. By contrast, MR is constant (and equal to the market equilibrium price *P*) for competitive firms, so that they do not have to worry about MR per unit falling as they produce more units. As a result, competitive firms increase production by a larger amount than imperfectly competitive firms whenever resource prices fall.

FIGURE 14.2 The imperfectly competitive seller's demand curve for a resource. An imperfectly competitive seller's resource demand curve *D* (solid) slopes downward because both marginal product and product price fall as resource employment and output rise. This downward slope is greater than that for a purely competitive seller (dashed resource demand curve) because the pure competitor can sell the added output at a constant price.



WORKED PROBLEMS

W14.1
Resource demand

Market Demand for a Resource

The total, or market, demand curve for a specific resource shows the various total amounts of the resource that firms will purchase or hire at various resource prices, other things equal. Recall that the total, or market, demand curve for a *product* is found by summing horizontally the demand curves of all individual buyers in the market. The market demand curve for a particular *resource* is derived in essentially the same way—by summing horizontally the individual demand or MRP curves for all firms hiring that resource.

QUICK REVIEW 14.1

- To maximize profit, a firm will purchase or hire a resource in an amount at which the resource's marginal revenue product equals its marginal resource cost (MRP = MRC).
- Application of the MRP = MRC rule to a firm's MRP curve demonstrates that the MRP curve is the firm's resource demand curve. In a purely competitive resource market, resource price (the wage rate) equals MRC.
- The resource demand curve of a purely competitive seller is downsloping solely because the marginal product of the resource diminishes; the resource demand curve of an imperfectly competitive seller is downsloping because marginal product diminishes *and* product price falls as output is increased.

CONSIDER THIS . . .

Superstars

In what economist Robert Frank calls "winner-take-all markets," a few highly talented performers have huge earnings relative to the average performers in the market. Because consumers and firms seek out "top" performers, small differences in talent or popularity get magnified into huge differences in pay.

In these markets, consumer spending gets channeled toward a few performers. The media then "hypes" these individuals, which further increases the public's awareness of their talents. Many more consumers then buy the stars' products. Although it is not easy to stay on top, several superstars emerge.

The high earnings of superstars result from the high revenues they generate from their work. Consider Beyoncé Knowles. If she sold only a few thousand songs and attracted only a few hundred fans to each concert, the revenue she would produce—her marginal revenue product—would be quite modest. So, too, would be her earnings.

But consumers have anointed Beyoncé as queen of the R&B and hip-hop portion of pop culture. The demand for her music and concerts is extraordinarily high. She sells *millions* of songs, not thousands, and draws *thousands* to her concerts, not hundreds. Her extraordinarily high net earnings derive from her extraordinarily high MRP.

So it is for the other superstars in the "winner-take-all markets." Influenced by the media, but coerced by no one, consumers direct their spending toward a select few. The resulting strong demand for these stars' services reflects their high MRP. And because top talent (by definition) is very limited, superstars receive amazingly high earnings.

Determinants of Resource Demand

LO14.3 List the factors that increase or decrease resource demand.

What will alter the demand for a resource—that is, shift the resource demand curve? The fact that resource demand is derived from *product demand* and depends on *resource productivity* suggests two "resource demand shifters." Also, our analysis of how changes in the prices of other products can shift a product's demand curve (Chapter 3) suggests another factor: changes in the *prices of other resources*.

Changes in Product Demand

Other things equal, an increase in the demand for a product will increase the demand for a resource used in its production, whereas a decrease in product demand will decrease the demand for that resource.

Let's see how this works. The first thing to recall is that a change in the demand for a product will change its price. In Table 14.1, let's assume that an increase in product demand boosts product price from \$2 to \$3. You should calculate the new resource demand schedule (columns 1 and 6) that would result and plot it in Figure 14.1 to verify that the new resource demand curve lies to the right of the old demand curve. Similarly, a decline in the product demand (and price) will shift the resource demand curve to the left. This effect—resource demand changing along with product demand—demonstrates that resource demand is derived from product demand.

Example: Assuming no offsetting change in supply, a decrease in the demand for new houses will drive down house prices. Those lower prices will decrease the MRP of construction workers, and therefore the demand for construction workers will fall. The resource demand curve such as in Figure 14.1 or Figure 14.2 will shift to the left.

Changes in Productivity

Other things equal, an increase in the productivity of a resource will increase the demand for the resource and a decrease in productivity will reduce the demand for the resource. If we doubled the MP data of column 3 in

Table 14.1, the MRP data of column 6 would also double, indicating a rightward shift of the resource demand curve.

The productivity of any resource may be altered over the long run in several ways:

- ***Quantities of other resources*** The marginal productivity of any resource will vary with the quantities of the other resources used with it. The greater the amount of capital and land resources used with, say, labor, the greater will be labor's marginal productivity and, thus, labor demand.
- ***Technological advance*** Technological improvements that increase the quality of other resources, such as capital, have the same effect. The better the *quality* of capital, the greater the productivity of labor used with it. Dockworkers employed with a specific amount of real capital in the form of unloading cranes are more productive than dockworkers with the same amount of real capital embodied in older conveyor-belt systems.
- ***Quality of the variable resource*** Improvements in the quality of the variable resource, such as labor, will increase its marginal productivity and therefore its demand. In effect, there will be a new demand curve for a different, more skilled, kind of labor.

All these considerations help explain why the average level of (real) wages is higher in industrially advanced nations (for example, the United States, Germany, Japan, and France) than in developing nations (for example, Nicaragua, Ethiopia, Angola, and Cambodia). Workers in industrially advanced nations are generally healthier, better educated, and better trained than are workers in developing countries. Also, in most industries they work with a larger and more efficient stock of capital goods and more abundant natural resources. This increases productivity and creates a strong demand for labor. On the supply side of the market, labor is scarcer relative to capital in industrially advanced than in most developing nations. A strong demand and a relatively scarce supply of labor result in high wage rates in the industrially advanced nations.

Changes in the Prices of Other Resources

Changes in the prices of other resources may change the demand for a specific resource. For example, a change in the price of capital may change the demand for labor. The direction of the change in labor demand will depend on whether labor and capital are substitutes or complements in production.

Substitute Resources Suppose the technology in a certain production process is such that labor and capital are substitutable. A firm can produce some specific amount of output using a relatively small amount of labor and a relatively large amount of capital, or vice versa. Now assume that the price of machinery (capital) falls. The effect on the demand for labor will be the net result of two opposed effects: the substitution effect and the output effect.

- ***Substitution effect*** The decline in the price of machinery prompts the firm to substitute machinery for labor. This allows the firm to produce its output at lower cost. So at the fixed wage rate, smaller quantities of labor are now employed. This **substitution effect** decreases the demand for labor. More generally, the substitution effect indicates that a firm will purchase more of an input whose relative price has declined and, conversely, use less of an input whose relative price has increased.
- ***Output effect*** Because the price of machinery has fallen, the costs of producing various outputs must also decline. With lower costs, the firm finds it profitable to produce and sell a greater output. The greater output increases the demand for all resources, including labor. So this **output effect** increases the demand for labor. More generally, the output effect means that the firm will purchase more of one particular input when the price of the other input falls and less of that particular input when the price of the other input rises.
- ***Net effect*** The substitution and output effects are both present when the price of an input changes, but they work in opposite directions. For a decline in the price of capital, the substitution effect decreases the demand for labor and the output effect increases it. The net change in labor demand depends on the relative sizes of the two effects: If the substitution effect outweighs the output effect, a decrease in the price of capital decreases the demand for labor. If the output effect exceeds the substitution effect, a decrease in the price of capital increases the demand for labor.

Complementary Resources Recall from Chapter 3 that certain products, such as computers and software, are complementary goods; they "go together" and are jointly demanded. Resources may also be complementary; an increase in the quantity of one of them used in the production process requires an increase in the amount used of the other as well, and vice versa. Suppose a small design firm does computer-assisted design (CAD) with relatively expensive personal computers as its basic piece of capital equipment. Each computer requires exactly one design engineer to

TABLE 14.3 **The Effect of an Increase in the Price of Capital on the Demand for Labor, D_L**

(1) Relationship of Inputs	(2) Increase in the Price of Capital (a) Substitution Effect	(b) Output Effect	(c) Combined Effect
Substitutes in production	Labor substituted for capital	Production costs up, output down, and less of both capital and labor used	D_L increases if the substitution effect exceeds the output effect; D_L decreases if the output effect exceeds the substitution effect
Complements in production	No substitution of labor for capital	Production costs up, output down, and less of both capital and labor used	D_L decreases (because only the output effect applies)

operate it; the machine is not automated—it will not run itself—and a second engineer would have nothing to do.

Now assume that a technological advance in the production of these computers substantially reduces their price. There can be no substitution effect because labor and capital must be used in *fixed proportions*, one person for one machine. Capital cannot be substituted for labor. But there *is* an output effect. Other things equal, the reduction in the price of capital goods means lower production costs. Producing a larger output will therefore be profitable. In doing so, the firm will use both more capital and more labor. When labor and capital are complementary, a decline in the price of capital increases the demand for labor through the output effect.

We have cast our analysis of substitute resources and complementary resources mainly in terms of a decline in the price of capital. Table 14.3 summarizes the effects of an *increase* in the price of capital on the demand for labor. Please study it carefully.

Now that we have discussed the full list of the determinants of labor demand, let's again review their effects. Stated in terms of the labor resource, the demand for labor will increase (the labor demand curve will shift rightward) when:

- The demand for (and therefore the price of) the product produced by that labor *increases*.
- The productivity (MP) of labor *increases*.
- The price of a substitute input *decreases*, provided the output effect exceeds the substitution effect.
- The price of a substitute input *increases*, provided the substitution effect exceeds the output effect.
- The price of a complementary input *decreases*.

Be sure that you can "reverse" these effects to explain a *decrease* in labor demand.

Table 14.4 provides several illustrations of the determinants of labor demand, listed by the categories of determinants we have discussed. You will benefit by giving them a close look.

Occupational Employment Trends

Changes in labor demand have considerable significance since they affect wage rates and employment in specific occupations. Increases in labor demand for certain occupational groups result in increases in their employment; decreases in labor demand result in decreases in their employment. For illustration, let's first look at occupations for which labor demand is growing and then examine occupations for which it is declining. (Wage rates are the subject of the next chapter.)

TABLE 14.4 **Determinants of Labor Demand: Factors That Shift the Labor Demand Curve**

Determinant	Examples
Change in product demand	Gambling increases in popularity, increasing the demand for workers at casinos. Consumers decrease their demand for leather coats, decreasing the demand for tanners. The federal government increases spending on homeland security, increasing the demand for security personnel.
Change in productivity	An increase in the skill levels of physicians increases the demand for their services. Computer-assisted graphic design increases the productivity of, and demand for, graphic artists.
Change in the price of another resource	An increase in the price of electricity increases the cost of producing aluminum and reduces the demand for aluminum workers. The price of security equipment used by businesses to protect against illegal entry falls, decreasing the demand for night guards. The price of cell phone equipment decreases, reducing the cost of cell phone service; this in turn increases the demand for cell phone assemblers. Health-insurance premiums rise, and firms substitute part-time workers who are not covered by insurance for full-time workers who are.

TABLE 14.5 **The 10 Fastest-Growing U.S. Occupations in Percentage Terms, 2010–2020**

Occupation	Employment, Thousands of Jobs 2010	2020	Percentage Increase*
Personal care aides	861	1,468	70.5
Home health aides	1,018	1,724	69.4
Biomedical engineers	16	25	61.7
Masonry helpers	29	47	60.1
Carpentry helpers	47	72	55.7
Veterinary technologists and technicians	80	122	52.0
Iron and rebar workers	19	28	48.6
Physical therapist assistants	67	98	45.7
Piping and plumbing helpers	58	84	45.4
Meeting, convention, and event planners	72	103	43.7

*Percentages and employment numbers may not reconcile due to rounding.
Source: Bureau of Labor Statistics, "Employment Projections," **www.bls.gov**.

TABLE 14.6 **The 10 Most Rapidly Declining U.S. Occupations in Percentage Terms, 2010–2020**

Occupation	Employment, Thousands of Jobs 2010	2020	Percentage Decrease*
Shoe machine operators	3	2	53.4
Postal service mail sorters	142	73	48.5
Postal service clerks	66	34	48.2
Fabric/apparel pattern makers	6	4	35.6
Postmasters/mail superintendents	25	18	27.8
Sewing machine operators	163	121	25.8
Switchboard operators	143	110	23.3
Textile cutting machine operators	15	12	21.8
Textile knitting/weaving machine operators	23	18	18.2
Semiconductor processors	21	17	17.9

*Percentages and employment numbers may not reconcile due to rounding.
Source: Bureau of Labor Statistics, "Employment Projections," **www.bls.gov**.

The Fastest-Growing Occupations Table 14.5 lists the 10 fastest-growing U.S. occupations for 2010 to 2020, as measured by percentage changes and projected by the Bureau of Labor Statistics. It is no coincidence that the service occupations dominate the list. In general, the demand for service workers in the United States is rapidly outpacing the demand for manufacturing, construction, and mining workers.

Of the 10 fastest-growing occupations in percentage terms, three—personal care aides (people who provide home health for the elderly and disabled), home health aides (people who provide short-term medical care after discharge from hospitals), and physical therapist assistants—are related to health care. The rising demands for these types of labor are derived from the growing demand for health services, caused by several factors. The aging of the U.S. population has brought with it more medical problems, the rising standard of income has led to greater expenditures on health care, and the continued presence of private and public insurance has allowed people to buy more health care than most could afford individually.

The Most Rapidly Declining Occupations In contrast, Table 14.6 lists the 10 U.S. occupations with the greatest projected job loss (in percentage terms) between 2010 and 2020. Several of the occupations owe their declines mainly to "labor-saving" technological change. For example, automated or computerized equipment has greatly reduced the need for postal employees, sewing machine operators, and pattern makers.

Five of the occupations in the declining employment list are related to textiles and apparel. The U.S. demand for these goods is increasingly being filled through imports. Those jobs are therefore rapidly disappearing in the United States.

As we indicated, the "top-10" lists shown in Tables 14.5 and 14.6 are based on percentage changes. In terms of absolute job growth and loss, the greatest projected employment growth between 2010 and 2020 is for home health aides (706,000 jobs) and personal care aides (607,000 jobs). The greatest projected absolute decline in employment is for postal service mail sorters (−71,000 jobs).

Elasticity of Resource Demand

LO14.4 Discuss the determinants of elasticity of resource demand.

The employment changes we have just discussed have resulted from shifts in the locations of resource demand curves. Such changes in demand must be distinguished from changes in the quantity of a resource demanded caused by a change in the price of the specific resource under consideration. Such a change is caused not by a shift of the demand curve but, rather, by a movement from one point to another on a fixed resource demand curve. Example: In Figure 14.1 we note that an increase in the wage rate from \$5 to \$7 will reduce the quantity of labor demanded from 5 to 4 units. This is a change in the *quantity of labor demanded* as distinct from a *change in the demand for labor*.

The sensitivity of resource quantity to changes in resource prices along a fixed resource demand curve is measured by the **elasticity of resource demand.** In coefficient form,

$$E_{rd} = \frac{\text{percentage change in resource quantity demanded}}{\text{percentage change in resource price}}$$

ORIGIN OF THE IDEA

O14.1

Elasticity of resource demand

When E_{rd} is greater than 1, resource demand is elastic; when E_{rd} is less than 1, resource demand is inelastic; and when E_{rd} equals 1, resource demand is unit-elastic. What determines the elasticity of resource demand? Several factors are at work.

Ease of Resource Substitutability The degree to which resources are substitutable is a fundamental determinant of elasticity. More specifically, the greater the substitutability of other resources, the more elastic is the demand for a particular resource. As an example, the high degree to which computerized voice recognition systems are substitutable for human beings implies that the demand for human beings answering phone calls at call centers is quite elastic. In contrast, good substitutes for physicians are rare, so demand for them is less elastic or even inelastic. If a furniture manufacturer finds that several types of wood are equally satisfactory in making coffee tables, a rise in the price of any one type of wood may cause a sharp drop in the amount demanded as the producer substitutes some other type of wood for the type of wood whose price has gone up. At the other extreme, there may be no reasonable substitutes; bauxite is absolutely essential in the production of aluminum ingots. Thus, the demand for bauxite by aluminum producers is inelastic.

Time can play a role in the ease of input substitution. For example, a firm's truck drivers may obtain a substantial wage increase with little or no immediate decline in employment. But over time, as the firm's trucks wear out and are replaced, that wage increase may motivate the company to purchase larger trucks and in that way deliver the same total output with fewer drivers.

Elasticity of Product Demand Because the demand for labor is a derived demand, the elasticity of the demand for the output that the labor is producing will influence the elasticity of the demand for labor. Other things equal, the greater the price elasticity of product demand, the greater the elasticity of resource demand. For example, suppose that the wage rate falls. This means a decline in the cost of producing the product and a drop in the product's price. If the elasticity of product demand is great, the resulting increase in the quantity of the product demanded will be large and thus necessitate a large increase in the quantity of labor to produce the additional output. This implies an elastic demand for labor. But if the demand for the product is inelastic, the increase in the amount of the product demanded will be small, as will be the increases in the quantity of labor demanded. This suggests an inelastic demand for labor.

Remember that the resource demand curve in Figure 14.1 is more elastic than the resource demand curve shown in Figure 14.2. The difference arises because in Figure 14.1 we assume a perfectly elastic product demand curve, whereas Figure 14.2 is based on a downsloping or less than perfectly elastic product demand curve.

Ratio of Resource Cost to Total Cost The larger the proportion of total production costs accounted for by a resource, the greater the elasticity of demand for that resource. In the extreme, if labor cost is the only production cost, then a 20 percent increase in wage rates will shift all the firm's cost curves upward by 20 percent. If product demand is elastic, this substantial increase in costs will cause a relatively large decline in sales and a sharp decline in the amount of labor demanded. So labor demand is highly elastic. But if labor cost is only 50 percent of production cost, then a 20 percent increase in wage rates will increase costs by only 10 percent. With the same elasticity of product demand, this will cause a relatively small decline in sales and therefore in the amount of labor demanded. In this case the demand for labor is much less elastic.

QUICK REVIEW 14.2

- A resource demand curve will shift because of changes in product demand, changes in the productivity of the resource, and changes in the prices of other inputs.
- If resources A and B are substitutable, a decline in the price of A will decrease the demand for B provided the substitution effect exceeds the output effect. But if the output effect exceeds the substitution effect, the demand for B will increase.
- If resources C and D are complements, a decline in the price of C will increase the demand for D.
- Elasticity of resource demand measures the extent to which producers change the quantity of a resource they hire when its price changes.
- For any particular resource, the elasticity of resource demand will be less the greater the difficulty of substituting other resources for the resource, the smaller the elasticity of product demand, and the smaller the proportion of total cost accounted for by the resource.

Optimal Combination of Resources*

LO14.5 Determine how a competitive firm selects its optimal combination of resources.

So far, our main focus has been on one variable input, labor. But in the long run firms can vary the amounts of all the resources they use. That's why we need to consider what combination of resources a firm will choose when *all* its inputs are variable. While our analysis is based on two resources, it can be extended to any number of inputs.

We will consider two interrelated questions:

- What combination of resources will minimize costs at a specific level of output?
- What combination of resources will maximize profit?

The Least-Cost Rule

A firm is producing a specific output with the **least-cost combination of resources** when the last dollar spent on each resource yields the same marginal product. That is, the cost of any output is minimized when the ratios of marginal product to price of the last units of resources used are the same for each resource. To see how this rule maximizes profits in a more concrete setting, consider firms that are competitive buyers in resource markets. Because each firm is too small to affect resource prices, each firm's marginal resource costs will equal market resource prices and each firm will be able to hire as many or as few units as it would like of any and all resources at their respective market prices. Thus, if there are just two resources, labor and capital, a competitive firm will minimize its total cost of a specific output when

$$\frac{\text{Marginal product of labor } (MP_L)}{\text{Price of labor } (P_L)} = \frac{\text{Marginal product of capital } (MP_C)}{\text{Price of capital } (P_C)} \qquad (1)$$

Throughout, we will refer to the marginal products of labor and capital as MP_L and MP_C, respectively, and symbolize the price of labor by P_L and the price of capital by P_C.

A concrete example will show why fulfilling the condition in equation 1 leads to least-cost production. Assume that the price of both capital and labor is $1 per unit but that Siam Soups currently employs them in such amounts that the marginal product of labor is 10 and the marginal product of capital is 5. Our equation immediately tells us that this is not the least costly combination of resources:

$$\frac{MP_L = 10}{P_L = \$1} > \frac{MP_C = 5}{P_C = \$1}$$

Suppose Siam spends $1 less on capital and shifts that dollar to labor. It loses 5 units of output produced by the last dollar's worth of capital, but it gains 10 units of output from the extra dollar's worth of labor. Net output increases by 5 (= 10 − 5) units for the same total cost. More such shifting of dollars from capital to labor will push the firm *down* along its MP curve for labor and *up* along its MP curve for capital, increasing output and moving the firm toward a position of equilibrium where equation 1 is fulfilled. At that equilibrium position, the MP per dollar for the last unit of both labor and capital might be, for example, 7. And Siam will be producing a greater output for the same (original) cost.

Whenever the same total-resource cost can result in a greater total output, the cost per unit—and therefore the total cost of any specific level of output—can be reduced. Being able to produce a *larger* output with a *specific* total cost is the same as being able to produce a *specific* output with a *smaller* total cost. If Siam buys $1 less of capital, its output will fall by 5 units. If it spends only $.50 of that dollar on labor, the firm will increase its output by a compensating 5 units (= $\frac{1}{2}$ of the MP per dollar). Then the firm will realize the same total output at a $0.50 lower total cost.

The cost of producing any specific output can be reduced as long as equation 1 does not hold. But when dollars have been shifted between capital and labor to the point where equation 1 holds, no additional changes in the use of capital and labor will reduce costs further. Siam will be producing that output using the least-cost combination of capital and labor.

All the long-run cost curves developed in Chapter 9 and used thereafter assume that the least-cost combination of inputs has been realized at each level of output. Any firm that combines resources in violation of the least-cost rule would have a higher-than-necessary average total cost at each level of output. That is, it would incur *X-inefficiency*, as discussed in Figure 12.7.

The producer's least-cost rule is analogous to the consumer's utility-maximizing rule described in Chapter 7. In achieving the utility-maximizing combination of goods, the consumer considers both his or her preferences as reflected in diminishing-marginal-utility data and the prices of the various products. Similarly, in achieving the cost-minimizing combination of resources, the producer considers both the marginal-product data and the prices (costs) of the various resources.

*Note to Instructors: We consider this section to be optional. If desired, it can be skipped without loss of continuity. It can also be deferred until after the discussion of wage determination in the next chapter.

The Profit-Maximizing Rule

Minimizing cost is not sufficient for maximizing profit. A firm can produce any level of output in the least costly way by applying equation 1. But only one unique level of output maximizes profit. Our earlier analysis of product markets showed that this profit-maximizing output occurs where marginal revenue equals marginal cost (MR = MC). Near the beginning of this chapter we determined that we could write this profit-maximizing condition as MRP = MRC as it relates to resource inputs.

In a purely competitive resource market the marginal resource cost (MRC) is equal to the resource price P. Thus, for any competitive resource market, we have as our profit-maximizing equation

$$\text{MRP (resource)} = P\text{ (resource)}$$

This condition must hold for every variable resource, and in the long run all resources are variable. In competitive markets, a firm will therefore achieve its **profit-maximizing combination of resources** when each resource is employed to the point at which its marginal revenue product equals its resource price. For two resources, labor and capital, we need both

$$P_L = \text{MRP}_L \qquad \text{and} \qquad P_C = \text{MRP}_C$$

We can combine these conditions by dividing both sides of each equation by their respective prices and equating the results to get

$$\frac{\text{MRP}_L}{P_L} = \frac{\text{MRP}_C}{P_C} = 1 \qquad (2)$$

Note in equation 2 that it is not sufficient that the MRPs of the two resources be *proportionate* to their prices; the MRPs must be *equal* to their prices and the ratios therefore equal to 1. For example, if $\text{MRP}_L = \$15$, $P_L = \$5$, $\text{MRP}_C = \$9$, and $P_C = \$3$, Siam is underemploying both capital and labor even though the ratios of MRP to resource price are identical for both resources. The firm can expand its profit by hiring additional amounts of both capital and labor until it moves down its downsloping MRP curves to the points at which $\text{MRP}_L = \$5$ and $\text{MRP}_C = \$3$. The ratios will then be 5/5 and 3/3 and equal to 1.

The profit-maximizing position in equation 2 includes the cost-minimizing condition of equation 1. That is, if a firm is maximizing profit according to equation 2, then it must be using the least-cost combination of inputs to do so. However, the converse is not true: A firm operating at least cost according to equation 1 may not be operating at the output that maximizes its profit.

WORKED PROBLEMS

W14.2
Optimal combination of resources

Numerical Illustration

A numerical illustration will help you understand the least-cost and profit-maximizing rules. In columns 2, 3, 2′, and 3′ in Table 14.7 we show the total products and marginal products for various amounts of labor and capital that are assumed to be the only inputs Siam needs in producing its soup. Both inputs are subject to diminishing returns.

We also assume that labor and capital are supplied in competitive resource markets at \$8 and \$12, respectively, and that Siam's soup sells competitively at \$2 per unit. For

TABLE 14.7 Data for Finding the Least-Cost and Profit-Maximizing Combination of Labor and Capital, Siam's Soups*

Labor (Price = \$8)					Capital (Price = \$12)				
(1) Quantity	(2) Total Product (Output)	(3) Marginal Product	(4) Total Revenue	(5) Marginal Revenue Product	(1′) Quantity	(2′) Total Product (Output)	(3′) Marginal Product	(4′) Total Revenue	(5′) Marginal Revenue Product
0	0		\$ 0		0	0		\$ 0	
1	12	12	24	\$24	1	13	13	26	\$26
2	22	10	44	20	2	22	9	44	18
3	28	6	56	12	**3**	28	6	56	**12**
4	33	5	66	10	4	32	4	64	8
5	37	4	74	**8**	5	35	3	70	6
6	40	3	80	6	6	37	2	74	4
7	42	2	84	4	7	38	1	76	2

*To simplify, it is assumed in this table that the productivity of each resource is independent of the quantity of the other. For example, the total and marginal products of labor are assumed not to vary with the quantity of capital employed.

both labor and capital we can determine the total revenue associated with each input level by multiplying total product by the $2 product price. These data are shown in columns 4 and 4′. They enable us to calculate the marginal revenue product of each successive input of labor and capital as shown in columns 5 and 5′, respectively.

Producing at Least Cost What is the least-cost combination of labor and capital for Siam to use in producing, say, 50 units of output? The answer, which we can obtain by trial and error, is 3 units of labor and 2 units of capital. Columns 2 and 2′ indicate that this combination of labor and capital does, indeed, result in the required 50 (= 28 + 22) units of output. Now, note from columns 3 and 3′ that hiring 3 units of labor gives us $MP_L/P_L = \frac{6}{8} = \frac{3}{4}$ and hiring 2 units of capital gives us $MP_C/P_C = \frac{9}{12} = \frac{3}{4}$. So equation 1 is fulfilled. How can we verify that costs are actually minimized? First, we see that the total cost of employing 3 units of labor and 2 of capital is $48 [= (3 × $8) + (2 × $12)].

Other combinations of labor and capital will also yield 50 units of output, but at a higher cost than $48. For example, 5 units of labor and 1 unit of capital will produce 50 (= 37 + 13) units, but total cost is higher, at $52 [= (5 × $8) + (1 × $12)]. This comes as no surprise because 5 units of labor and 1 unit of capital violate the least-cost rule—$MP_L/P_L = \frac{4}{8}$, $MP_C/P_C = \frac{13}{12}$. Only the combination (3 units of labor and 2 units of capital) that minimizes total cost will satisfy equation 1. All other combinations capable of producing 50 units of output violate the cost-minimizing rule, and therefore cost more than $48.

Maximizing Profit Will 50 units of output maximize Siam's profit? No, because the profit-maximizing terms of equation 2 are not satisfied when the firm employs 3 units of labor and 2 of capital. To maximize profit, each input should be employed until its price equals its marginal revenue product. But for 3 units of labor, labor's MRP in column 5 is $12 while its price is only $8. This means the firm could increase its profit by hiring more labor. Similarly, for 2 units of capital, we see in column 5′ that capital's MRP is $18 and its price is only $12. This indicates that more capital should also be employed. By producing only 50 units of output (even though they are produced at least cost), labor and capital are being used in less-than-profit-maximizing amounts. The firm needs to expand its employment of labor and capital, thereby increasing its output.

Table 14.7 shows that the MRPs of labor and capital are equal to their prices, so equation 2 is fulfilled when Siam is employing 5 units of labor and 3 units of capital. So this is the profit-maximizing combination of inputs.[2] The firm's total cost will be $76, made up of $40 (= 5 × $8) of labor and $36 (= 3 × $12) of capital. Total revenue will be $130, found either by multiplying the total output of 65 (= 37 + 28) by the $2 product price or by summing the total revenues attributable to labor ($74) and to capital ($56). The difference between total revenue and total cost in this instance is $54 (= $130 − $76). Experiment with other combinations of labor and capital to demonstrate that they yield an economic profit of less than $54.

Note that the profit-maximizing combination of 5 units of labor and 3 units of capital is also a least-cost combination for this particular level of output. Using these resource amounts satisfies the least-cost requirement of equation 1 in that $MP_L/P_L = \frac{4}{8} = \frac{1}{2}$ and $MP_C/P_C = \frac{6}{12} = \frac{1}{2}$.

Marginal Productivity Theory of Income Distribution

LO14.6 Explain the marginal productivity theory of income distribution.

Our discussion of resource pricing is the cornerstone of the controversial view that fairness and economic justice are one of the outcomes of a competitive capitalist economy. Table 14.7 demonstrates, in effect, that workers receive income payments (wages) equal to the marginal contributions they make to their employers' outputs and revenues. In other words, workers are paid according to the value of the labor services that they contribute to production. Similarly, owners of the other resources receive income based on the value of the resources they supply in the production process.

ORIGIN OF THE IDEA

O14.2

Marginal productivity theory of distribution

In this **marginal productivity theory of income distribution,** income is distributed according to contribution to society's output. So, if you are willing to accept the proposition "To each according to the value of what he or she creates," income payments based on marginal revenue product provide a fair and equitable distribution of society's income.

---

[2]Because we are dealing with discrete (nonfractional) units of the two outputs here, the use of 4 units of labor and 2 units of capital is equally profitable. The fifth unit of labor's MRP and its price (cost) are equal at $8, so that the fifth labor unit neither adds to nor subtracts from the firm's profit; similarly, the third unit of capital has no effect on profit.

LAST WORD

Input Substitution: The Case of ATMs

Banks Are Using More Automatic Teller Machines (ATMs) and Employing Fewer Human Tellers.

As you have learned from this chapter, a firm achieves its least-cost combination of inputs when the last dollar it spends on each input makes the same contribution to total output. This raises an interesting real-world question: What happens when technological advance makes available a new, highly productive capital good for which *MP*/*P* is greater than it is for other inputs, say, a particular type of labor? The answer is that the least-cost mix of resources abruptly changes, and the firm responds accordingly. If the new capital is a substitute for labor (rather than a complement), the firm replaces the particular type of labor with the new capital. That is exactly what is happening in the banking industry, in which ATMs are replacing human bank tellers.

ATMs made their debut at a bank in London in 1967. Shortly thereafter, U.S. firms Docutel and Diebold each introduced their own models. Today, Diebold and NCR (also a U.S. firm) dominate global sales, with the Japanese firm Fujitsu being a distant third. The number of ATMs and their usage have exploded, and currently there are nearly 400,000 ATMs in the United States. In 1975, about 10 *million* ATM transactions occurred in the United States. Today there are about 80 *billion* U.S. ATM transactions each year.

ATMs are highly productive: A single machine can handle hundreds of transactions daily, thousands weekly, and millions over the course of several years. ATMs can not only handle cash withdrawals but also accept deposits and facilitate switches of funds between various accounts. Although ATMs are expensive for banks to buy and install, they are available 24 hours a day, and their cost per transaction is one-fourth the cost for human tellers. They rarely get "held up," and they do not quit their jobs (turnover among human tellers is nearly 50 percent per year). Moreover, ATMs are highly convenient; unlike human tellers, they are located not only at banks but also at busy street corners, workplaces, universities, and malls. The same bank card that enables you to withdraw cash from a local ATM also enables you to withdraw pounds from an ATM in London, yen from an ATM in Tokyo, and rubles from an ATM in Moscow. (All this, of course, assumes that you have money in your account!)

In the terminology of this chapter, the more productive, lower-priced ATMs have reduced the demand for a substitute in production—human tellers. Between 1990 and 2000, an estimated 80,000 human teller positions were eliminated, and more positions may disappear in coming years. Where will the people holding these jobs go? Most will eventually move to other occupations. Although the lives of individual tellers are disrupted, society clearly wins. Society obtains more convenient banking services as well as the other goods that these "freed-up" labor resources help produce.

Source: Based partly on Ben Craig, "Where Have All the Tellers Gone?" Federal Reserve Bank of Cleveland, *Economic Commentary*, Apr. 15, 1997; and statistics provided by the American Bankers Association.

This sounds reasonable, but you need to be aware of serious criticisms of this theory of income distribution:

- ***Inequality*** Critics argue that the distribution of income resulting from payment according to marginal productivity may be highly unequal because productive resources are very unequally distributed in the first place. Aside from their differences in mental and physical attributes, individuals encounter substantially different opportunities to enhance their productivity through education and training and the use of more and better equipment. Some people may not be able to participate in production at all because of mental or physical disabilities, and they would obtain no income under a system of distribution based solely on marginal productivity. Ownership of property resources is also highly unequal. Many owners of land and capital resources obtain their property by inheritance rather than through their own productive effort. Hence, income from inherited property resources conflicts with the "To each according to the value of what he or she creates" idea. Critics say that these inequalities call for progressive taxation and government spending programs aimed at creating an income distribution that will be more equitable than that which would occur if the income distribution were made strictly according to marginal productivity.

- ***Market imperfections*** The marginal productivity theory of income distribution rests on the assumptions of competitive markets. But, as we will see in Chapter 15, not all labor markets are highly competitive. In some labor markets employers exert their wage-setting power to pay less-than-competitive wages. And some workers, through labor unions, professional associations, and occupational licensing laws, wield wage-setting power in selling their services. Even the process of collective bargaining over wages suggests a power struggle over the division of income. In wage setting through negotiations, market forces—and income shares based on marginal productivity—may get partially pushed into the background. In addition, discrimination in the labor market can distort earnings patterns. In short, because of real-world market imperfections, wage rates and other resource prices are not always based solely on contributions to output.

QUICK REVIEW 14.3

- Any specific level of output will be produced with the least-costly combination of variable resources when the marginal product per dollar's worth of each input is the same.
- A firm is employing the profit-maximizing combination of resources when each resource is used to the point where its marginal revenue product equals its price.
- The marginal productivity theory of income distribution holds that all resources are paid according to their marginal contributions to output.

SUMMARY

LO14.1 Explain the significance of resource pricing.

Resource prices help determine money incomes, and they simultaneously ration resources to various industries and firms.

LO14.2 Convey how the marginal revenue productivity of a resource relates to a firm's demand for that resource.

The demand for any resource is derived from the product it helps produce. That means the demand for a resource will depend on its productivity and on the market value (price) of the good it is used to produce.

Marginal revenue product is the extra revenue a firm obtains when it employs 1 more unit of a resource. The marginal revenue product curve for any resource is the demand curve for that resource because the firm equates resource price and MRP in determining its profit-maximizing level of resource employment. Thus each point on the MRP curve indicates how many resource units the firm will hire at a specific resource price.

The firm's demand curve for a resource slopes downward because the marginal product of additional units declines in accordance with the law of diminishing returns. When a firm is selling in an imperfectly competitive market, the resource demand curve falls for a second reason: Product price must be reduced for the firm to sell a larger output. The market demand curve for a resource is derived by summing horizontally the demand curves of all the firms hiring that resource.

LO14.3 List the factors that increase or decrease resource demand.

The demand curve for a resource will shift as the result of (*a*) a change in the demand for, and therefore the price of, the product the resource is producing; (*b*) changes in the productivity of the resource; and (*c*) changes in the prices of other resources.

If resources A and B are substitutable for each other, a decline in the price of A will decrease the demand for B provided the substitution effect is greater than the output effect. But if the output effect *exceeds* the substitution effect, a decline in the price of A will increase the demand for B.

If resources C and D are complementary or jointly demanded, there is only an output effect; a change in the price of C will change the demand for D in the opposite direction.

The majority of the 10 fastest-growing occupations in the United States—by percentage increase—relate to health care and computers (review Table 14.5); the 10 most rapidly declining occupations by percentage decrease, however, are more mixed (review Table 14.6).

LO14.4 Discuss the determinants of elasticity of resource demand.

The elasticity of demand for a resource measures the responsiveness of producers to a change in the resource's price. The coefficient of the elasticity of resource demand is

$$E_{rd} = \frac{\text{percentage change in resource quantity demanded}}{\text{percentage change in resource price}}$$

When E_{rd} is greater than 1, resource demand is elastic; when E_{rd} is less than 1, resource demand is inelastic; and when E_{rd} equals 1, resource demand is unit-elastic.

The elasticity of demand for a resource will be greater (*a*) the greater the ease of substituting other resources for labor, (*b*) the greater the elasticity of demand for the product, and (*c*) the larger the proportion of total production costs attributable to the resource.

LO14.5 Determine how a competitive firm selects its optimal combination of resources.

Any specific level of output will be produced with the least costly combination of variable resources when the marginal product per dollar's worth of each input is the same—that is, when

$$\frac{\text{MP of labor}}{\text{Price of labor}} = \frac{\text{MP of capital}}{\text{Price of capital}}$$

A firm is employing the profit-maximizing combination of resources when each resource is used to the point where its marginal revenue product equals its price. In terms of labor and capital, that occurs when the MRP of labor equals the price of labor and the MRP of capital equals the price of capital—that is, when

$$\frac{\text{MRP of labor}}{\text{Price of labor}} = \frac{\text{MRP of capital}}{\text{Price of capital}} = 1$$

LO14.6 Explain the marginal productivity theory of income distribution.

The marginal productivity theory of income distribution holds that resources are paid according to their marginal contribution to output. Critics say that such an income distribution is too unequal and that real-world market imperfections result in pay above and below marginal contributions to output.

TERMS AND CONCEPTS

derived demand
marginal product (MP)
marginal revenue product (MRP)
marginal resource cost (MRC)
MRP = MRC rule
substitution effect
output effect
elasticity of resource demand
least-cost combination of resources
profit-maximizing combination of resources
marginal productivity theory of income distribution

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What is the significance of resource pricing? Explain how the factors determining resource demand differ from those determining product demand. Explain the meaning and significance of the fact that the demand for a resource is a derived demand. Why do resource demand curves slope downward? **LO14.1**
2. In 2009 General Motors (GM) announced that it would reduce employment by 21,000 workers. What does this decision reveal about how GM viewed its marginal revenue product (MRP) and marginal resource cost (MRC)? Why didn't GM reduce employment by more than 21,000 workers? By fewer than 21,000 workers? **LO14.3**
3. What factors determine the elasticity of resource demand? What effect will each of the following have on the elasticity or the location of the demand for resource C, which is being used to produce commodity X? Where there is any uncertainty as to the outcome, specify the causes of that uncertainty. **LO14.4**
   a. An increase in the demand for product X.
   b. An increase in the price of substitute resource D.
   c. An increase in the number of resources substitutable for C in producing X.
   d. A technological improvement in the capital equipment with which resource C is combined.
   e. A fall in the price of complementary resource E.
   f. A decline in the elasticity of demand for product X due to a decline in the competitiveness of product market X.
4. In each of the following four cases, MRP_L and MRP_C refer to the marginal revenue products of labor and capital, respectively, and P_L and P_C refer to their prices. Indicate in each case whether the conditions are consistent with maximum profits for the firm. If not, state which resource(s) should be used in larger amounts and which resource(s) should be used in smaller amounts. **LO14.5**
   a. $MRP_L = \$8$; $P_L = \$4$; $MRP_C = \$8$; $P_C = \$4$.
   b. $MRP_L = \$10$; $P_L = \$12$; $MRP_C = \$14$; $P_C = \$9$.
   c. $MRP_L = \$6$; $P_L = \$6$; $MRP_C = \$12$; $P_C = \$12$.
   d. $MRP_L = \$22$; $P_L = \$26$; $MRP_C = \$16$; $P_C = \$19$.
5. Florida citrus growers say that the recent crackdown on illegal immigration is increasing the market wage rates necessary

to get their oranges picked. Some are turning to $100,000 to $300,000 mechanical harvesters known as “trunk, shake, and catch” pickers, which vigorously shake oranges from the trees. If widely adopted, what will be the effect on the demand for human orange pickers? What does that imply about the relative strengths of the substitution and output effects? **LO14.5**

6. **LAST WORD** Explain the economics of the substitution of ATMs for human tellers. Some banks are beginning to assess transaction fees when customers use human tellers rather than ATMs. What are these banks trying to accomplish?

REVIEW QUESTIONS

1. Cindy is a baker and runs a large cupcake shop. She has already hired 11 employees and is thinking of hiring a 12th. Cindy estimates that a 12th worker would cost her $100 per day in wages and benefits while increasing her total revenue from $2,600 per day to $2,750 per day. Should Cindy hire a 12th worker? **LO14.2**
   - **a.** Yes.
   - **b.** No.
   - **c.** You need more information to figure this out.
2. Complete the following labor demand table for a firm that is hiring labor competitively and selling its product in a competitive market. **LO14.2**

Units of Labor	Total Product	Marginal Product	Product Price	Total Revenue	Marginal Revenue Product
0	0		$2	$______	
1	17	______	2	______	$______
2	31	______	2	______	______
3	43	______	2	______	______
4	53	______	2	______	______
5	60	______	2	______	______
6	65	______	2	______	______

   - **a.** How many workers will the firm hire if the market wage rate is $27.95? $19.95? Explain why the firm will not hire a larger or smaller number of units of labor at each of these wage rates.
   - **b.** Show in schedule form and graphically the labor demand curve of this firm.
   - **c.** Now again determine the firm’s demand curve for labor, assuming that it is selling in an imperfectly competitive market and that, although it can sell 17 units at $2.20 per unit, it must lower product price by 5 cents in order to sell the marginal product of each successive labor unit. Compare this demand curve with that derived in part *b*. Which curve is more elastic? Explain.
3. Alice runs a shoemaking factory that utilizes both labor and capital to make shoes. Which of the following would shift the factory’s demand for capital? You can select one or more answers from the choices shown. **LO14.3**
   - **a.** Many consumers decide to walk barefoot all the time.
   - **b.** New shoemaking machines are twice as efficient as older machines.
   - **c.** The wages that the factory has to pay its workers rise due to an economy-wide labor shortage.
4. FreshLeaf is a commercial salad maker that produces “salad in a bag” that is sold at many local supermarkets. Its customers like lettuce but don’t care so much what type of lettuce is included in each bag of salad, so you would expect FreshLeaf’s demand for iceberg lettuce to be: **LO14.4**
   - **a.** Elastic.
   - **b.** Inelastic.
   - **c.** Unit elastic.
   - **d.** All of the above.
5. Suppose the productivity of capital and labor are as shown in the table below. The output of these resources sells in a purely competitive market for $1 per unit. Both capital and labor are hired under purely competitive conditions at $3 and $1, respectively. **LO14.5**
   - **a.** What is the least-cost combination of labor and capital the firm should employ in producing 80 units of output? Explain.
   - **b.** What is the profit-maximizing combination of labor and capital the firm should use? Explain. What is the resulting level of output? What is the economic profit? Is this the least costly way of producing the profit-maximizing output?

Units of Capital	MP of Capital	Units of Labor	MP of Labor
0		0	
	24		11
1		1	
	21		9
2		2	
	18		8
3		3	
	15		7
4		4	
	9		6
5		5	
	6		4
6		6	
	3		1
7		7	
	1		$\frac{1}{2}$
8		8	

6. A software company in Silicon Valley uses programmers (labor) and computers (capital) to produce apps for mobile devices. The firm estimates that when it comes to labor, MP_L = 5 apps per month while P_L = $1,000 per month. And when it comes to capital, MP_C = 8 apps per month

while P_C = $1,000 per month. If the company wants to maximize its profits, it should: LO14.5

a. Increase labor while decreasing capital.
b. Decrease labor while increasing capital.
c. Keep the current amounts of capital and labor just as they are.
d. None of the above.

PROBLEMS

1. A delivery company is considering adding another vehicle to its delivery fleet; each vehicle is rented for $100 per day. Assume that the additional vehicle would be capable of delivering 1,500 packages per day and that each package that is delivered brings in ten cents in revenue. Also assume that adding the delivery vehicle would not affect any other costs. LO14.2
   a. What is the MRP? What is the MRC? Should the firm add this delivery vehicle?
   b. Now suppose that the cost of renting a vehicle doubles to $200 per day. What are the MRP and MRC? Should the firm add a delivery vehicle under these circumstances?
   c. Next suppose that the cost of renting a vehicle falls back down to $100 per day but, due to extremely congested freeways, an additional vehicle would only be able to deliver 750 packages per day. What are the MRP and MRC in this situation? Would adding a vehicle under these circumstances increase the firm's profits?

2. Suppose that marginal product tripled while product price fell by one-half in Table 14.1. What would be the new MRP values in Table 14.1? What would be the net impact on the location of the resource demand curve in Figure 14.1? LO14.2

3. Suppose that a monopoly firm finds that its MR is $50 for the first unit sold each day, $49 for the second unit sold each day, $48 for the third unit sold each day, and so on. Further suppose that the first worker hired produces 5 units per day, the second 4 units per day, the third 3 units per day, and so on. LO14.3
   a. What is the firm's MRP for each of the first five workers?
   b. Suppose that the monopolist is subjected to rate regulation and the regulator stipulates that it must charge exactly $40 per unit for all units sold. At that price, what is the firm's MRP for each of the first five workers?
   c. If the daily wage paid to workers is $170 per day, how many workers will the unregulated monopoly demand? How many will the regulated monopoly demand? Looking at those figures, will the regulated or the unregulated monopoly demand more workers at that wage?
   d. If the daily wage paid to workers falls to $77 per day, how many workers will the unregulated monopoly demand? How many will the regulated monopoly demand? Looking at those figures, will the regulated or the unregulated monopoly demand more workers at that wage?
   e. Comparing your answers to parts *c* and *d*, does regulating a monopoly's output price *always* increase its demand for resources?

4. Consider a small landscaping company run by Mr. Viemeister. He is considering increasing his firm's capacity. If he adds one more worker, the firm's total monthly revenue will increase from $50,000 to $58,000. If he adds one more tractor, monthly revenue will increase from $50,000 to $62,000. Each additional worker costs $4,000 per month, while an additional tractor would also cost $4,000 per month. LO14.5
   a. What is the marginal product of labor? The marginal product of capital?
   b. What is the ratio of the marginal product of labor to the price of labor (MP_L/P_L)? What is the ratio of the marginal product of capital to the price of capital (MP_K/P_K)?
   c. Is the firm using the least-costly combination of inputs?
   d. Does adding an additional worker or adding an additional tractor yield a larger increase in total revenue for each dollar spent?

CHAPTER **15**

Wage Determination

Learning Objectives

LO15.1 Explain why labor productivity and real hourly compensation track so closely over time.

LO15.2 Show how wage rates and employment levels are determined in competitive labor markets.

LO15.3 Demonstrate how monopsony (a market with a single employer) can reduce wages below competitive levels.

LO15.4 Discuss how unions increase wage rates by pursuing the demand-enhancement model, the craft union model, or the industrial union model.

LO15.5 Explain why wages and employment are determined by collective bargaining in a situation of bilateral monopoly.

LO15.6 Discuss how minimum wage laws affect labor markets.

LO15.7 List the major causes of wage differentials.

LO15.8 Identify the types, benefits, and costs of "pay-for-performance" plans.

LO15.9 (Appendix) Relate who belongs to U.S. unions, the basics of collective bargaining, and the economic effects of unions.

Nearly 140 million Americans go to work each day. We work at an amazing variety of jobs for thousands of different firms and receive considerable differences in pay. What determines our hourly wage or annual salary? Why is the salary for, say, a topflight major-league baseball player $15 million or more a year, whereas the pay for a first-rate schoolteacher is $50,000? Why are starting salaries for college graduates who major in engineering and accounting so much higher than those for graduates majoring in journalism and sociology?

Having explored the major factors that underlie labor demand, we now bring *labor supply* into our analysis to help answer these questions. Generally speaking, labor supply and labor demand interact to determine the level of hourly wage rates or annual salaries in each occupation. Collectively, those wages and salaries make up about 70 percent of all income paid to American resource suppliers.

Labor, Wages, and Earnings

LO15.1 Explain why labor productivity and real hourly compensation track so closely over time.

Economists use the term "labor" broadly to apply to (1) blue- and white-collar workers of all varieties; (2) professional people such as lawyers, physicians, dentists, and teachers; and (3) owners of small businesses, including barbers, plumbers, and a host of retailers who provide labor as they operate their own businesses.

Wages are the price that employers pay for labor. Wages not only take the form of direct money payments such as hourly pay, annual salaries, bonuses, commissions, and royalties but also fringe benefits such as paid vacations, health insurance, and pensions. Unless stated otherwise, we will use the term "wages" to mean all such payments and benefits converted to an hourly basis. That will remind us that the **wage rate** is the price paid per unit of labor services, in this case an hour of work. It will also let us distinguish between the wage rate and labor earnings, the latter determined by multiplying the number of hours worked by the hourly wage rate.

We must also distinguish between nominal wages and real wages. A **nominal wage** is the amount of money received per hour, day, or year. A **real wage** is the quantity of goods and services a worker can obtain with nominal wages; real wages reveal the "purchasing power" of nominal wages.

Your real wage depends on your nominal wage and the prices of the goods and services you purchase. Suppose you receive a 5 percent increase in your nominal wage during a certain year but in that same year the price level increases by 3 percent. Then your real wage has increased by 2 percent (= 5 percent − 3 percent). Unless otherwise indicated, we will assume that the overall level of prices remains constant. In other words, we will discuss only *real* wages.

General Level of Wages

Wages differ among nations, regions, occupations, and individuals. Wage rates are much higher in the United States than in China or India. They are slightly higher in the north and east of the United States than in the south. Plumbers are paid less than NFL punters. And one physician may earn twice as much as another physician for the same number of hours of work. The average wages earned by workers also differ by gender, race, and ethnic background.

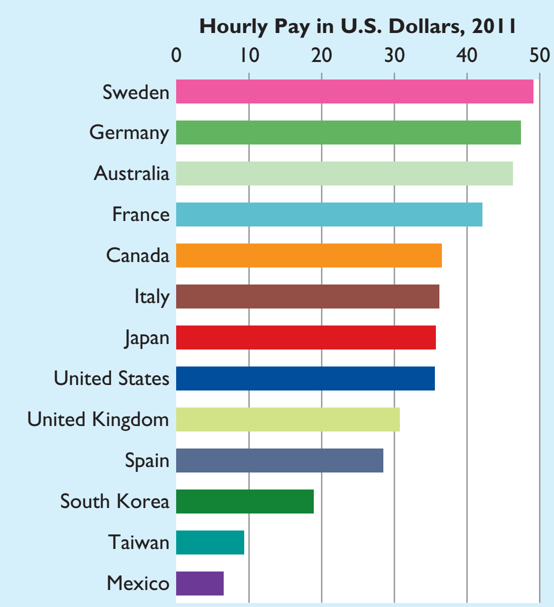
The general, or average, level of wages, like the general level of prices, includes a wide range of different wage rates. It includes the wages of bakers, barbers, brick masons, and brain surgeons. By averaging such wages, we can more easily compare wages among regions and among nations.

As Global Perspective 15.1 suggests, the general level of real wages in the United States is relatively high—although clearly not the highest in the world.

GLOBAL PERSPECTIVE 15.1

Hourly Wages of Production Workers, Selected Nations

Wage differences are pronounced worldwide. The data shown here indicate that hourly compensation in the United States is not as high as in some European nations. It is important to note, however, that the prices of goods and services vary greatly among nations and the process of converting foreign wages into dollars may not accurately reflect such variations.



Source: U.S. Bureau of Labor Statistics, **www.bls.gov**.

The simplest explanation for the high real wages in the United States and other industrially advanced economies (referred to hereafter as advanced economies) is that the demand for labor in those nations is relatively large compared to the supply of labor.

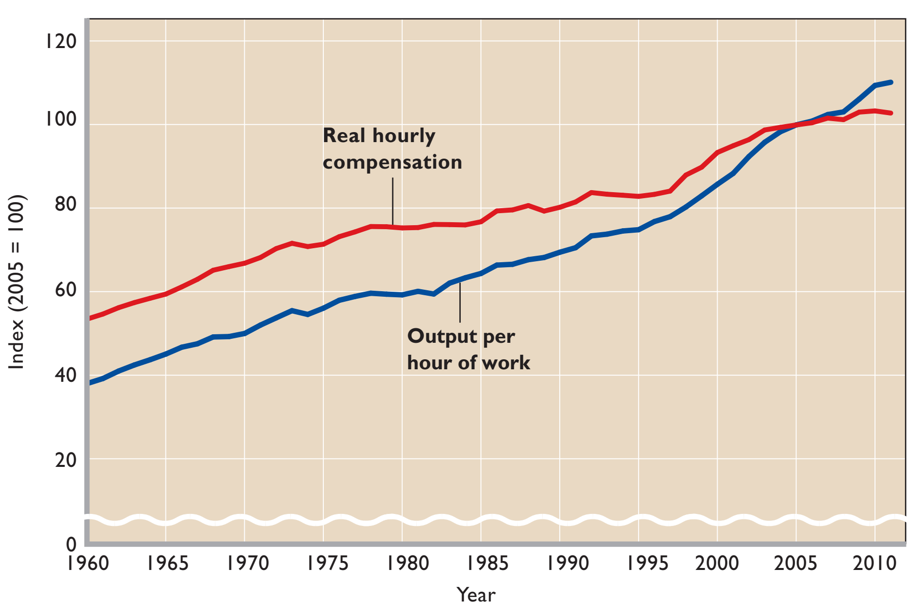
Role of Productivity

We know from the previous chapter that the demand for labor, or for any other resource, depends on its productivity. In general, the greater the productivity of labor, the greater is the demand for it. And if the total supply of labor is fixed, then the stronger the demand for labor, the higher is the average level of real wages. The demand for labor in the United States and the other major advanced economies is large because labor in those countries is highly productive. There are several reasons for that high productivity:

- ***Plentiful capital*** Workers in the advanced economies have access to large amounts of physical capital equipment (machinery and buildings). In the United States in 2011, $126,062 of physical capital was available, on average, for each worker.
- ***Access to abundant natural resources*** In advanced economies, natural resources tend to be abundant in relation to the size of the labor force. Some of those resources are available domestically and others are imported from abroad. The United States, for example, is richly endowed with arable land, mineral resources, and sources of energy for industry.
- ***Advanced technology*** The level of production technology is generally high in advanced economies. Not only do workers in these economies have more capital equipment to work with, but that equipment is technologically superior to the equipment available to the vast majority of workers worldwide. Moreover, work methods in the advanced economies are steadily being improved through scientific study and research.
- ***Labor quality*** The health, vigor, education, and training of workers in advanced economies are generally superior to those in developing nations. This means that, even with the same quantity and quality of natural and capital resources, workers in advanced economies tend to be more efficient than many of their foreign counterparts.
- ***Other factors*** Less obvious factors also may underlie the high productivity in some of the advanced economies. In the United States, for example, such factors include (a) the efficiency and flexibility of management; (b) a business, social, and political environment that emphasizes production and productivity; (c) the vast size of the domestic market, which enables firms to engage in mass production; and (d) the increased specialization of production enabled by free-trade agreements with other nations.

Real Wages and Productivity

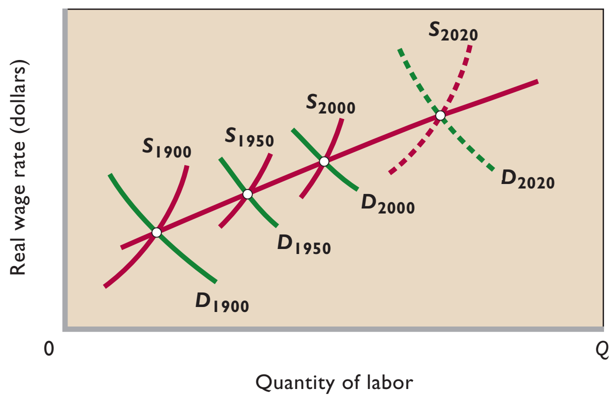
Figure 15.1 shows the close long-run relationship in the United States between output per hour of work and real hourly compensation (= wages and salaries + employers' contributions to social insurance and private benefit plans). Because real income and real output are two ways



Source: Bureau of Labor Statistics, **www.bls.gov**.

FIGURE 15.1 Output per hour and real hourly compensation in the United States, 1960–2011. Over long time periods, output per hour of work and real hourly compensation are closely related.

FIGURE 15.2 The long-run trend of real wages in the United States. The productivity of U.S. labor has increased substantially over the long run, causing the demand for labor *D* to shift rightward (that is, to increase) more rapidly than increases in the supply of labor *S*. The result has been increases in real wages.



of viewing the same thing, real income (compensation) per worker can increase only at about the same rate as output per worker. When workers produce more real output per hour, more real income is available to distribute to them for each hour worked.

In the actual economy, however, suppliers of land, capital, and entrepreneurial talent also share in the income from production. Real wages therefore do not always rise in lockstep with gains in productivity over short spans of time. But over long periods, productivity and real wages tend to rise together.

Long-Run Trend of Real Wages

Basic supply and demand analysis helps explain the long-term trend of real-wage growth in the United States. The nation's labor force has grown significantly over the decades. But, as a result of the productivity-increasing factors we have mentioned, increases in labor demand have outstripped increases in labor supply. Figure 15.2 shows several such increases in labor supply and labor demand. The result has been a long-run, or secular, increase in wage rates and employment. For example, real hourly compensation in the United States has roughly doubled since 1960. Over that same period, employment has increased by about 80 million workers.

A Purely Competitive Labor Market

LO15.2 Show how wage rates and employment levels are determined in competitive labor markets.

Average levels of wages, however, disguise the great variation of wage rates among occupations and within occupations. What determines the wage rate paid for a specific type of labor? Demand and supply analysis again is revealing. Let's begin by examining labor demand and labor supply in a **purely competitive labor market.** In this type of market:

- Numerous firms compete with one another in hiring a specific type of labor.
- Each of many qualified workers with identical skills supplies that type of labor.
- Individual firms and individual workers are "wage takers" since neither can exert any control over the market wage rate.

Market Demand for Labor

Suppose 200 firms demand a particular type of labor, say, carpenters. These firms need not be in the same industry; industries are defined according to the products they produce and not the resources they employ. Thus, firms producing wood-framed furniture, wood windows and doors, houses and apartment buildings, and wood cabinets will demand carpenters. To find the total, or market, labor demand curve for a particular labor service, we sum horizontally the labor demand curves (the marginal revenue product curves) of the individual firms, as indicated in **Figure 15.3 (Key Graph).** The horizontal summing of the 200 labor demand curves like *d* in Figure 15.3b yields the market labor demand curve *D* in Figure 15.3a.

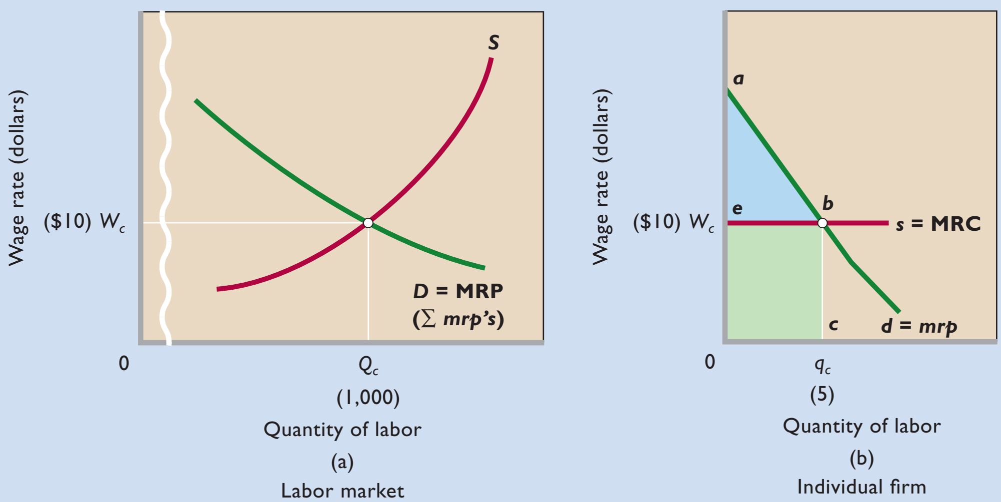
Market Supply of Labor

On the supply side of a purely competitive labor market, we assume that no union is present and that workers individually compete for available jobs. The supply curve for each type of labor slopes upward, indicating that employers as a group must pay higher wage rates to obtain more workers. They must do this to bid workers away from other industries, occupations, and localities. Within limits, workers have alternative job opportunities. For example, they may work in other industries in the same locality, or they may work in their present occupations in different cities or states, or they may work in other occupations.

Firms that want to hire these workers (here, carpenters) must pay higher wage rates to attract them away from the alternative job opportunities available to them. They must also pay higher wages to induce people who are not currently in the labor force—who are perhaps doing household activities or enjoying leisure—to seek employment. In short, assuming that wages are constant in other labor markets, higher wages in a particular labor market entice more workers to offer their labor services in that market—a fact expressed graphically by the upsloping market supply-of-labor curve *S* in Figure 15.3a.

KEY GRAPH

FIGURE 15.3 Labor supply and labor demand in (a) a purely competitive labor market and (b) a single competitive firm. In a purely competitive labor market (a), market labor supply *S* and market labor demand *D* determine the equilibrium wage rate W_c and the equilibrium number of workers Q_c. Each individual competitive firm (b) takes this competitive wage W_c as given. Thus, the individual firm's labor supply curve *s* = MRC is perfectly elastic at the going wage W_c. Its labor demand curve, *d*, is its MRP curve (here labeled *mrp*). The firm maximizes its profit by hiring workers up to where MRP = MRC. Area 0*abc* represents both the firm's total revenue and its total cost. The green area is its total wage cost; the blue area is its nonlabor costs, including a normal profit—that is, the firm's payments to the suppliers of land, capital, and entrepreneurship.



QUICK QUIZ FOR FIGURE 15.3

1. The supply-of-labor curve *S* slopes upward in graph (a) because:
   - **a.** the law of diminishing marginal utility applies.
   - **b.** the law of diminishing returns applies.
   - **c.** workers can afford to "buy" more leisure when the wage rate increases.
   - **d.** higher wages are needed to attract workers away from other labor markets, household activities, and leisure.
2. This firm's labor demand curve *d* in graph (b) slopes downward because:
   - **a.** the law of diminishing marginal utility applies.
   - **b.** the law of diminishing returns applies.
   - **c.** the firm must lower its price to sell additional units of its product.
   - **d.** the firm is a competitive employer, not a monopsonist.
3. In employing five workers, the firm represented in graph (b):
   - **a.** has a total wage cost of $6,000.
   - **b.** is adhering to the general principle of undertaking all actions for which the marginal benefit exceeds the marginal cost.
   - **c.** uses less labor than would be ideal from society's perspective.
   - **d.** experiences increasing marginal returns.
4. A rightward shift of the labor supply curve in graph (a) would shift curve:
   - **a.** *d* = *mrp* leftward in graph (b).
   - **b.** *d* = *mrp* rightward in graph (b).
   - **c.** *s* = MRC upward in graph (b).
   - **d.** *s* = MRC downward in graph (b).

Answers: 1. d; 2. b; 3. b; 4. d

Labor Market Equilibrium

The intersection of the market labor demand curve and the market labor supply curve determines the equilibrium wage rate and level of employment in a purely competitive labor market. In Figure 15.3a the equilibrium wage rate is W_c ($10) and the number of workers hired is Q_c (1,000). To the individual firm the market wage rate W_c is given. Each of the many firms employs such a small fraction of the total available supply of this type of labor that no single firm can influence the wage rate. As shown by the horizontal line *s* in Figure 15.3b, the supply of labor faced by an individual firm is perfectly elastic. It can hire as many or as few workers as it wants to at the market wage rate.

TABLE 15.1 The Supply of Labor: Pure Competition in the Hire of Labor

(1) Units of Labor	(2) Wage Rate	(3) Total Labor Cost	(4) Marginal Resource (Labor) Cost
0	$10	$ 0	
1	10	10	$10
2	10	20	10
3	10	30	10
4	10	40	10
5	10	50	10
6	10	60	10

Each individual firm will maximize its profit (or minimize its loss) by hiring this type of labor up to the point at which marginal revenue product is equal to marginal resource cost. This is merely an application of the MRP = MRC rule we developed in Chapter 14.

As Table 15.1 indicates, when an individual competitive firm faces the market price for a resource, the marginal cost of that resource (MRC) is constant and is equal to the market price for each and every unit that the competitive firm may choose to purchase. Note that MRC is constant at $10 and matches the $10 wage rate. Each additional worker hired adds precisely his or her own wage rate ($10 in this case) to the firm's total resource cost. So the firm in a purely competitive labor market maximizes its profit by hiring workers up to the point at which its wage rate equals MRP. In Figure 15.3b this firm will hire q_c (5) workers, paying each worker the market wage rate W_c ($10). The other 199 firms (not shown) that are hiring workers in this labor market will also each employ 5 workers and pay $10 per hour.

To determine a firm's total revenue from employing a particular number of labor units, we sum the MRPs of those units. For example, if a firm employs 3 labor units with marginal revenue products of $14, $13, and $12, respectively, then the firm's total revenue is $39 (= $14 + $13 + $12). In Figure 15.3b, where we are not restricted to whole units of labor, total revenue is represented by area 0*abc* under the MRP curve to the left of q_c. And what area represents the firm's total cost, including a normal profit? Answer: For q_c units, the same area—0*abc*. The green rectangle represents the firm's total wage cost ($0q_c \times 0W_c$). The blue triangle (total revenue minus total wage cost) represents the firm's nonlabor costs—its explicit and implicit payments to land, capital, and entrepreneurship. Thus, in this case, total cost (wages plus other income payments) equals total revenue. This firm and others like it are earning only a normal profit. So Figure 15.3b represents a long-run equilibrium for a firm that is selling its product in a purely competitive product market and hiring its labor in a purely competitive labor market.

CONSIDER THIS . . .

Fringe Benefits vs. Take-Home Pay

Figure 15.2 shows that total compensation has risen significantly over the past several decades. Not shown in that figure, however, is the fact that the amount of take-home pay received by middle-class American workers has increased by much less. One contributing factor has been the rise of fringe benefits.

To see why fringe benefits matter, recall that throughout this chapter we have defined the wage as the total price that employers pay to obtain labor and compensate workers for providing it. Under our definition, wages are the sum of take-home pay (such as hourly pay and annual salaries) and fringe benefits (such as paid vacations, health insurance, and pensions).

So now consider an equilibrium wage, such as W_c in Figure 15.3. If workers want higher fringe benefits, they can have them—but only if take-home pay falls by an equal amount. With the equilibrium wage fixed by supply and demand, the only way workers can get more fringe benefits is by accepting lower take-home pay.

This is an important point to understand because in recent decades, workers have received an increasing fraction of their total compensation in the form of fringe benefits—especially health insurance. Those fringe benefits are costly and in a competitive labor market, each $1 increase in fringe benefits means $1 less for paychecks.

That trade-off helps to explain why take-home pay has increased by less than total compensation in recent decades. With a rising fraction of total compensation flowing toward fringe benefits, the increase in take-home pay was much less than the overall increase in total compensation.

Monopsony Model

LO15.3 Demonstrate how monopsony (a market with a single employer) can reduce wages below competitive levels.

In the purely competitive labor market described in the preceding section, each employer hires too small an amount of labor to influence the wage rate. Each firm can hire as little or as much labor as it needs, but only at the market wage rate, as reflected in its horizontal labor supply curve. The situation is quite different when the labor market is a **monopsony,** a market structure in which there is only a single buyer. A labor market monopsony has the following characteristics:

- There is only a single buyer of a particular type of labor.
- The workers providing this type of labor have few employment options other than working for the monopsony because they are either geographically immobile or because finding alternative employment would mean having to acquire new skills.
- The firm is a "wage maker" because the wage rate it must pay varies directly with the number of workers it employs.

As is true of monopoly power, there are various degrees of monopsony power. In *pure* monopsony such power is at its maximum because only a single employer hires labor in the labor market. The best real-world examples are probably the labor markets in some towns that depend almost entirely on one major firm. For example, a silver-mining company may be almost the only source of employment in a remote Idaho town. A Colorado ski resort, a Wisconsin paper mill, or an Alaskan fish processor may provide most of the employment in its geographically isolated locale.

ORIGIN OF THE IDEA

O15.1

Monopsony

In other cases three or four firms may each hire a large portion of the supply of labor in a certain market and therefore have some monopsony power. Moreover, if they tacitly or openly act in concert in hiring labor, they greatly enhance their monopsony power.

Upsloping Labor Supply to Firm

When a firm hires most of the available supply of a certain type of labor, its decision to employ more or fewer workers affects the wage rate it pays to those workers. Specifically, if a firm is large in relation to the size of the labor market, it will have to pay a higher wage rate to attract labor away from other employment or from leisure. Suppose that there is only one employer of a particular type of labor in a certain geographic area. In this pure monopsony situation, the labor supply curve for the *firm* and the total labor supply curve for the *labor market* are identical. The monopsonist's supply curve—represented

FIGURE 15.4 The wage rate and level of employment in a monopsonistic labor market. In a monopsonistic labor market the employer's marginal resource (labor) cost curve (MRC) lies above the labor supply curve S. Equating MRC with MRP at point b, the monopsonist hires Q_m workers (compared with Q_c under competition). As indicated by point c on S, it pays only wage rate W_m (compared with the competitive wage W_c).

by curve S in Figure 15.4—is upsloping because the firm must pay higher wage rates if it wants to attract and hire additional workers. This same curve is also the monopsonist's average-cost-of-labor curve. Each point on curve S indicates the wage rate (cost) per worker that must be paid to attract the corresponding number of workers.

MRC Higher Than the Wage Rate

When a monopsonist pays a higher wage to attract an additional worker, it must pay that higher wage not only to the additional worker, but to all the workers it is currently employing at a lower wage. If not, labor morale will deteriorate, and the employer will be plagued with labor unrest because of wage-rate differences existing for the same job. Paying a uniform wage to all workers means that the cost of an extra worker—the marginal resource (labor) cost (MRC)—is the sum of that worker's wage rate and the amount necessary to bring the wage rate of all current workers up to the new wage level.

Table 15.2 illustrates this point. One worker can be hired at a wage rate of \$6. But hiring a second worker forces the firm to pay a higher wage rate of \$7. The marginal resource (labor) cost of the second worker is \$8—the \$7 paid to the second worker plus a \$1 raise for the first worker. From another viewpoint, total labor cost is now \$14 (= 2 × \$7), up from \$6 (= 1 × \$6). So the MRC of the second worker is \$8 (= \$14 − \$6), not just the \$7 wage rate paid to that worker. Similarly, the marginal labor cost of the third worker is \$10—the \$8 that must be paid to attract this worker from alternative employment plus \$1 raises, from \$7 to \$8, for the first two workers.

TABLE 15.2 The Supply of Labor: Monopsony in the Hiring of Labor

(1) Units of Labor	(2) Wage Rate	(3) Total Labor Cost	(4) Marginal Resource (Labor) Cost
0	$5	$ 0	
1	6	6	$ 6
2	7	14	8
3	8	24	10
4	9	36	12
5	10	50	14
6	11	66	16

Here is the key point: Because the monopsonist is the only employer in the labor market, its marginal resource (labor) cost exceeds the wage rate. Graphically, the monopsonist's MRC curve lies above the average-cost-of-labor curve, or labor supply curve S, as is clearly shown in Figure 15.4.

Equilibrium Wage and Employment

How many units of labor will the monopsonist hire, and what wage rate will it pay? To maximize profit, the monopsonist will employ the quantity of labor Q_m in Figure 15.4, because at that quantity MRC and MRP are equal (point b).[1] The monopsonist next determines how much it must pay to attract these Q_m workers. From the supply curve S, specifically point c, it sees that it must pay wage rate W_m. Clearly, it need not pay a wage equal to MRP; it can attract and hire exactly the number of workers it wants (Q_m) with wage rate W_m. And that is the wage that it will pay.

[1]The fact that MRC exceeds resource price when resources are hired or purchased under imperfectly competitive (monopsonistic) conditions calls for adjustments in Chapter 14's least-cost and profit-maximizing rules for hiring resources. (See equations 1 and 2 in the "Optimal Combination of Resources" section of Chapter 14.) Specifically, we must substitute MRC for resource price in the denominators of our two equations. That is, with imperfect competition in the hiring of both labor and capital, equation 1 becomes

$$\frac{MP_L}{MRC_L} = \frac{MP_C}{MRC_C} \qquad (1')$$

and equation 2 is restated as

$$\frac{MRP_L}{MRC_L} = \frac{MRP_C}{MRC_C} = 1 \qquad (2')$$

In fact, equations 1 and 2 can be regarded as special cases of 1′ and 2′ in which firms happen to be hiring under purely competitive conditions and resource price is therefore equal to, and can be substituted for, marginal resource cost.

Contrast these results with those that would prevail in a competitive labor market. With competition in the hiring of labor, the level of employment would be greater (at Q_c) and the wage rate would be higher (at W_c). Other things equal, the monopsonist maximizes its profit by hiring a smaller number of workers and thereby paying a less-than-competitive wage rate. Society obtains a smaller output, and workers receive a wage rate that is less by bc than their marginal revenue product. Just as a monopolistic seller finds it profitable to restrict product output to realize an above-competitive price for its goods, the monopsonistic employer of resources finds it profitable to restrict employment in order to reduce wage rates below those that would occur under competitive conditions.

WORKED PROBLEMS

W15.1
Labor markets: competition and monopsony

Examples of Monopsony Power

Fortunately, monopsonistic labor markets are uncommon in the United States. In most labor markets, several potential employers compete for most workers, particularly for workers who are occupationally and geographically mobile. Also, where monopsony labor market outcomes might have otherwise occurred, unions have often sprung up to counteract that power by forcing firms to negotiate wages. Nevertheless, economists have found some evidence of monopsony power in such diverse labor markets as the markets for nurses, professional athletes, public school teachers, newspaper employees, and some building-trade workers.

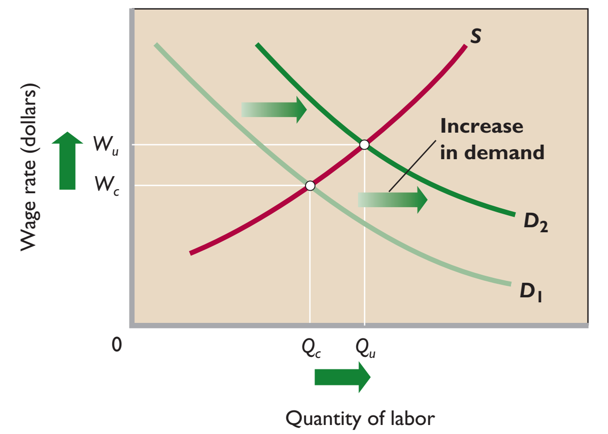
In the case of nurses, the major employers in most locales are a relatively small number of hospitals. Further, the highly specialized skills of nurses are not readily transferable to other occupations. It has been found, in accordance with the monopsony model, that, other things equal, the smaller the number of hospitals in a town or city (that is, the greater the degree of monopsony), the lower the beginning salaries of nurses.

Professional sports leagues also provide a good example of monopsony, particularly as it relates to the pay of first-year players. The National Football League, the National Basketball Association, and Major League Baseball assign first-year players to teams through "player drafts." That device prohibits other teams from competing for a player's services, at least for several years, until the player becomes a "free agent." In this way each league exercises monopsony power, which results in lower salaries than would occur under competitive conditions.

QUICK REVIEW 15.1

- Real wages have increased over time in the United States because labor demand has increased relative to labor supply.
- Over the long term, real wages per worker have increased at approximately the same rate as worker productivity.
- The competitive employer is a wage taker and employs workers at the point where the wage rate (= MRC) equals MRP.
- The labor supply curve to a monopsonist is upsloping, causing MRC to exceed the wage rate for each worker. Other things equal, the monopsonist, hiring where MRC = MRP, will employ fewer workers and pay a lower wage rate than would a purely competitive employer.

FIGURE 15.5 Unions and demand enhancement. When unions can increase the demand for union labor (say, from D_1 to D_2), they can realize higher wage rates (W_c to W_u) and more jobs (Q_c to Q_u).



Three Union Models

LO15.4 Discuss how unions increase wage rates by pursuing the demand-enhancement model, the craft union model, or the industrial union model.

Our assumption thus far has been that workers compete with one another in selling their labor services. But in some labor markets workers unionize and sell their labor services collectively. (We examine union membership, collective bargaining, and union impacts in detail in an appendix to this chapter. Here our focus is on three union wage models.)

When a union is formed in an otherwise competitive labor market, it usually bargains with a relatively large number of employers. It has many goals, the most important of which is to raise wage rates. It can pursue that objective in several ways.

Demand-Enhancement Model

Unions recognize that their ability to influence the demand for labor is limited. But, from the union's viewpoint, increasing the demand for union labor is highly desirable. As Figure 15.5 shows, an increase in the demand for union labor will create a higher union wage along with more jobs.

Unions can increase the demand for their labor by increasing the demand for the goods or services they help produce. Political lobbying is the main tool for increasing the demand for union-produced goods or services. For example, construction unions have lobbied for new highways, mass-transit systems, and stadium projects. Teachers' unions and associations have pushed for increased public spending on education. Unions in the aerospace industry have lobbied to increase spending on the military and on space exploration. U.S. steel unions and forest-product workers have lobbied for tariffs and quotas on foreign imports of steel and lumber, respectively. Such trade restrictions shift the demand for labor away from foreign countries and toward unionized U.S. labor.

Unions can also increase the demand for union labor by altering the price of other inputs. For example, although union members are generally paid significantly more than the minimum wage, unions have strongly supported increases in the minimum wage. The purpose may be to raise the price of low-wage, nonunion labor, which in some cases is substitutable for union labor. A higher minimum wage for nonunion workers will discourage employers from substituting such workers for union workers and will thereby bolster the demand for union members.

Similarly, unions have sometimes sought to increase the demand for their labor by supporting policies that will reduce or hold down the price of a complementary resource. For example, unions in industries that represent workers who transport fruits and vegetables may support legislation that allows low-wage foreign agricultural workers to temporarily work in the United States. Where union labor and another resource are complementary, a price decrease for the other resource will increase the demand for union labor through Chapter 14's output effect.

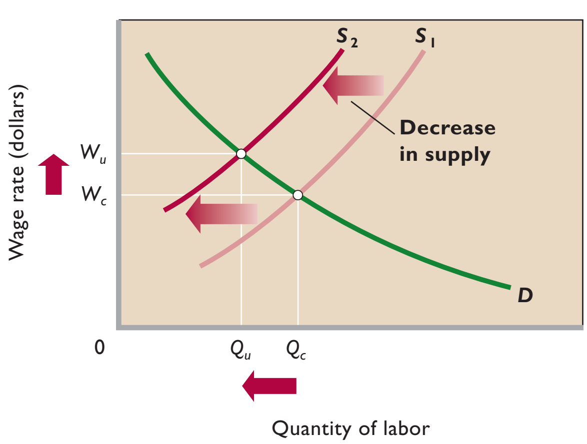
Exclusive or Craft Union Model

Unions can also boost wage rates by reducing the supply of labor, and over the years organized labor has favored policies to do just that. For example, labor unions have supported legislation that has (1) restricted permanent immigration, (2) reduced child labor, (3) encouraged compulsory retirement, and (4) enforced a shorter workweek.

Moreover, certain types of workers have adopted techniques designed to restrict the number of workers who can join their union. This is especially true of *craft unions*, whose members possess a particular skill, such as carpenters, brick masons, or plumbers. Craft unions have frequently forced employers to agree to hire only union members, thereby gaining virtually complete control of the labor supply. Then, by following restrictive membership policies—for example, long apprenticeships, very high initiation fees, and limits on the number of new members admitted—they have artificially restricted labor supply. As indicated in Figure 15.6, such practices result in higher wage rates and constitute what is called **exclusive unionism.** By excluding workers from unions and therefore from the labor supply, craft unions succeed in elevating wage rates.

This craft union model is also applicable to many professional organizations, such as the American Medical Association, the National Education Association, the American Bar Association, and hundreds of others. Such groups seek to prohibit competition for their services from less qualified labor suppliers. One way to accomplish that is through **occupational licensing.** Here a group of workers in a given occupation pressure federal, state, or municipal government to pass a law that says that some occupational group (for example, barbers, physicians, lawyers, plumbers, cosmetologists, egg graders, pest controllers) can practice their trade only if they meet certain requirements. Those requirements might include level of education, amount of work experience, the passing of an examination, and personal characteristics ("the practitioner must be of good moral character").

FIGURE 15.6 Exclusive or craft unionism. By reducing the supply of labor (say, from S_1 to S_2) through the use of restrictive membership policies, exclusive unions achieve higher wage rates (W_c to W_u). However, restriction of the labor supply also reduces the number of workers employed (Q_c to Q_u).



Members of the licensed occupation typically dominate the licensing board that administers such laws. The result is self-regulation, which often leads to policies that serve only to restrict entry to the occupation and reduce labor supply.

The expressed purpose of licensing is to protect consumers from incompetent practitioners—surely a worthy goal. But such licensing, if abused, results in above-competitive wages and earnings for those in the licensed occupation (Figure 15.6). Moreover, licensing requirements often include a residency requirement, which inhibits the interstate movement of qualified workers. Some 600 occupations are now licensed in the United States.

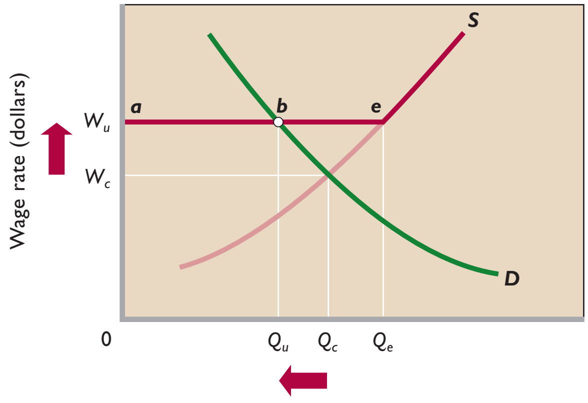
Inclusive or Industrial Union Model

Instead of trying to limit their membership, however, most unions seek to organize all available workers. This is especially true of the *industrial unions*, such as those of the automobile workers and steelworkers. Such unions seek as members all available unskilled, semiskilled, and skilled workers in an industry. It makes sense for a union to be exclusive when its members are skilled craft workers for whom the employer has few substitutes. But it does not make sense for a union to be exclusive when trying to organize unskilled and semiskilled workers. To break a strike, employers could then easily substitute unskilled or semiskilled nonunion workers for the unskilled or semiskilled union workers.

By contrast, an industrial union that includes virtually all available workers in its membership can put firms under great pressure to agree to its wage demands. Because of its legal right to strike, such a union can threaten to deprive firms of their entire labor supply. And an actual strike can do just that. Further, with virtually all available workers in the union, it will be difficult in the short run for new nonunion firms to emerge and thereby undermine what the union is demanding from existing firms.

We illustrate such **inclusive unionism** in Figure 15.7. Initially, the competitive equilibrium wage rate is W_c and the level of employment is Q_c. Now suppose an industrial union is formed that demands a higher, above-equilibrium wage rate of, say, W_u. That wage rate W_u would create a perfectly elastic labor supply over the range *ae* in Figure 15.7. If firms wanted to hire any workers in this range, they would have to pay the union-imposed wage rate. If they decide against meeting this wage demand, the union will supply no labor at all, and the firms will be faced with a strike. If firms decide it is better to pay the higher wage rate than to suffer a strike, they will cut back on employment from Q_c to Q_u.

FIGURE 15.7 Inclusive or industrial unionism. By organizing virtually all available workers in order to control the supply of labor, inclusive industrial unions may impose a wage rate, such as W_u, which is above the competitive wage rate W_c. In effect, this changes the labor supply curve from *S* to *aeS*. At wage rate W_u, employers will cut employment from Q_c to Q_u.



By agreeing to the union's wage demand, individual employers become wage takers at the union wage rate W_u. Because labor supply is perfectly elastic over range *ae*, the marginal resource (labor) cost is equal to the wage rate W_u over this range. The Q_u level of employment is the result of employers' equating this MRC (now equal to the union wage rate) with MRP, according to our profit-maximizing rule.

Note from point *e* on labor supply curve *S* that Q_e workers desire employment at wage W_u. But as indicated by point *b* on labor demand curve *D*, only Q_u workers are employed. The result is a surplus of labor of $Q_e - Q_u$ (also shown by distance *eb*). In a purely competitive labor market without the union, the effect of a surplus of unemployed workers would be lower wages. Specifically, the wage rate would fall to the equilibrium level W_c where the quantity of labor supplied equals the quantity of labor demanded (each Q_c). But this drop in wages does not happen because workers are acting collectively through their union. Individual workers cannot offer to work for less than W_u nor can employers pay less than that.

Wage Increases and Job Loss

Have U.S. unions been successful in raising the wages of their members? Evidence suggests that union members on average achieve a 15 percent wage advantage over nonunion workers. But when unions are successful in raising wages, their efforts also have another major effect.

As Figures 15.6 and 15.7 suggest, the wage-raising actions achieved by both exclusive and inclusive unionism reduce employment in unionized firms. Simply put, a union's success in achieving above-equilibrium wage rates tends to be accompanied by a decline in the number of workers employed. That result acts as a restraining influence on union wage demands. A union cannot expect to maintain solidarity within its ranks if it seeks a wage rate so high that 20 to 30 percent of its members lose their jobs.

Bilateral Monopoly Model

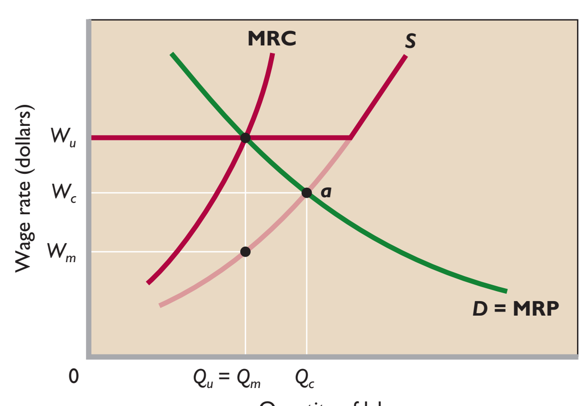
LO15.5 Explain why wages and employment are determined by collective bargaining in a situation of bilateral monopoly.

Suppose a strong industrial union is formed in a monopsonist labor market rather that a competitive labor market, thereby creating a combination of the monopsony model and the inclusive unionism model. Economists call the result **bilateral monopoly** because in its pure form there is a single seller and a single buyer. The union is a monopolistic "seller" of labor that controls labor supply and can influence wage rates, but it faces a monopsonistic "buyer" of labor that can also affect wages by altering the amount of labor that it employs. This is not an uncommon case, particularly in less pure forms in which a single union confronts two, three, or four large employers. Examples: steel, automobiles, construction equipment, professional sports, and commercial aircraft.

Indeterminate Outcome of Bilateral Monopoly

We show this situation in Figure 15.8, where Figure 15.7 is superimposed onto Figure 15.4. The monopsonistic employer will seek the below-competitive-equilibrium wage rate W_m, and the union will press for some

FIGURE 15.8 Bilateral monopoly in the labor market. A monopsonist seeks to hire Q_m workers (where MRC = MRP) and pay wage rate W_m corresponding to quantity Q_m on labor supply curve *S*. The inclusive union it faces seeks the above-equilibrium wage rate W_u. The actual outcome cannot be predicted by economic theory. It will result from bargaining between the two parties.



above-competitive-equilibrium wage rate such as W_u. Which will be the outcome? We cannot say with certainty. The outcome is "logically indeterminate" because the bilateral monopoly model does not explain what will happen at the bargaining table. We can expect the wage outcome to lie somewhere between W_m and W_u. Beyond that, about all we can say is that the party with the greater bargaining power and the more effective bargaining strategy will probably get a wage closer to the one it seeks.

Desirability of Bilateral Monopoly

The wage and employment outcomes in this situation might be more economically desirable than the term "bilateral monopoly" implies. The monopoly on one side of the market might in effect cancel out the monopoly on the other side, yielding competitive or near-competitive results. If either the union or management prevailed in this market—that is, if the actual wage rate were either W_u or W_m—employment would be restricted to Q_m (where MRP = MRC), which is below the competitive level.

But now suppose the monopoly power of the union roughly offsets the monopsony power of management, and the union and management agree on wage rate W_c, which is the competitive wage. Once management accepts this wage rate, its incentive to restrict employment disappears; no longer can it depress wage rates by restricting employment. Instead, management hires at the most profitable resource quantity, where the bargained wage rate W_c (which is now the firm's MRC) is equal to the MRP. It hires Q_c workers. Thus, with monopoly on both sides of the labor market, the resulting wage rate and level of employment may be closer to competitive levels than would be the case if monopoly existed on only one side of the market.

QUICK REVIEW 15.2

- In the demand-enhancement union model, a union increases the wage rate by increasing labor demand through actions that increase product demand or alter the prices of related inputs.
- In the exclusive (craft) union model, a union increases wage rates by artificially restricting labor supply, through, say, long apprenticeships or occupational licensing.
- In the inclusive (industrial) union model, a union raises the wage rate by gaining control over a firm's labor supply and threatening to withhold labor via a strike unless a negotiated wage is obtained.
- Bilateral monopoly occurs in a labor market where a monopsonist bargains with an inclusive, or industrial, union. Wage and employment outcomes are determined by collective bargaining in this situation.

The Minimum-Wage Controversy

LO15.6 Discuss how minimum wage laws affect labor markets.

Since the passage of the Fair Labor Standards Act in 1938, the United States has had a federal **minimum wage.** That wage has ranged between 30 and 50 percent of the average wage paid to manufacturing workers and was most recently raised to $7.25 in July 2009. Numerous states, however, have minimum wages that are higher than the federal minimum wage. Some of these state minimum wages are considerably higher. For example, in 2013 the minimum wage in the state of Washington was $9.19 an hour. The purpose of the minimum wage is to provide a "wage floor" that will help less-skilled workers earn enough income to escape poverty.

Case against the Minimum Wage

Critics, reasoning in terms of Figure 15.7, contend that an above-equilibrium minimum wage (say, W_u) will simply cause employers to hire fewer workers. Downsloping labor demand curves are a reality. The higher labor costs may even force some firms out of business. Then some of the poor, low-wage workers whom the minimum wage was designed to help will find themselves out of work. Critics point out that a worker who is *unemployed* and desperate to find a job at a minimum wage of $7.25 per hour is clearly worse off than he or she would be if *employed* at a market wage rate of, say, $6.50 per hour.

A second criticism of the minimum wage is that it is "poorly targeted" to reduce household poverty. Critics point out that much of the benefit of the minimum wage accrues to workers, including many teenagers, who do not live in impoverished households.

Case for the Minimum Wage

Advocates of the minimum wage say that critics analyze its impact in an unrealistic context. Figure 15.7, advocates claim, assumes a competitive labor market. But in a less competitive, low-pay labor market where employers possess some monopsony power (Figure 15.8), the minimum wage can increase wage rates without causing significant unemployment. Indeed, a higher minimum wage may even produce more jobs by eliminating the motive that monopsonistic firms have for restricting employment. For example, a minimum-wage floor of W_c in Figure 15.8 would change the firm's labor supply curve to W_caS and prompt the firm to increase its employment from Q_m workers to Q_c workers.

Moreover, even if the labor market is competitive, the higher wage rate might prompt firms to find more productive tasks for low-paid workers, thereby raising their productivity. Alternatively, the minimum wage may reduce *labor turnover* (the rate at which workers voluntarily quit). With fewer low-productive trainees, the *average* productivity of the firm's workers would rise. In either case, the alleged negative employment effects of the minimum wage might not occur.

Evidence and Conclusions

Which view is correct? Unfortunately, there is no clear answer. All economists agree that firms will not hire workers who cost more per hour than the value of their hourly output. So there is some minimum wage sufficiently high that it would severely reduce employment. Consider $30 an hour, as an absurd example. Because the majority of U.S. workers earned less than $20 per hour in 2011, a minimum wage of $30 per hour would render the majority of American workers unemployable because the minimum wage that they would have to be paid by potential employers would far exceed their marginal revenue products.

It has to be remembered, though, that a minimum wage will only cause unemployment in labor markets where the minimum wage is higher than the equilibrium wage. Because the current minimum wage of $7.25 per hour is much lower than the average hourly wage of about $19.78 that was earned by American workers in 2011, any unemployment caused by the $7.25 per hour minimum wage is most likely to fall on low-skilled workers who earn low wages due to their low productivity. These workers are mostly teenagers, adults who did not complete high school, and immigrants with low levels of education and poor English proficiency. For members of such groups, recent research suggests that a 10 percent increase in the minimum wage will cause a 1 to 3 percent decline in employment. However, estimates of the employment effect of minimum wage laws vary from study to study so that significant controversy remains.

The overall effect of the minimum wage is thus uncertain. On the one hand, the employment and unemployment effects of the minimum wage do not appear to be as great as many critics fear. On the other hand, because a large part of its effect is dissipated on nonpoverty families, the minimum wage is not as strong an antipoverty tool as many supporters contend.

Voting patterns and surveys make it clear, however, that the minimum wage has strong political support. Perhaps this stems from two realities: (1) More workers are believed to be helped than hurt by the minimum wage and (2) the minimum wage gives society some assurance that employers are not "taking undue advantage" of vulnerable, low-skilled workers.

TABLE 15.3 Average Annual Wages in Selected Occupations, 2011

Occupation	Average Annual Wages
1. Surgeons	$234,950
2. Petroleum engineers	138,980
3. Financial managers	120,450
4. Aircraft pilots	118,070
5. Law professors	108,760
6. Chemical engineers	99,440
7. Dental hygienists	69,760
8. Registered nurses	69,110
9. Police officers	56,260
10. Electricians	52,910
11. Carpenters	44,330
12. Travel agents	35,740
13. Barbers	28,050
14. Retail salespersons	25,020
15. Janitors	24,840
16. Child care workers	21,320
17. Fast food cooks	18,720

Source: Bureau of Labor Statistics, **www.bls.gov**.

Wage Differentials

LO15.7 List the major causes of wage differentials.

Hourly wage rates and annual salaries differ greatly among occupations. In Table 15.3 we list average annual salaries for a number of occupations to illustrate such occupational **wage differentials.** For example, observe that surgeons on average earn nine times as much as retail salespersons. Not shown, there are also large wage differentials within some of the occupations listed. For example, some highly experienced surgeons earn several times as much income as surgeons just starting their careers. And, although average wages for retail salespersons are relatively low, some top salespersons selling on commission make several times the average wages listed for their occupation.

What explains wage differentials such as these? Once again, the forces of demand and supply are revealing. As we demonstrate in Figure 15.9, wage differentials can arise on either the supply or the demand side of labor markets. Figure 15.9a and 15.9b represent labor markets for two occupational groups that have identical *labor supply curves*. Labor market (a) has a relatively high equilibrium wage (W_a) because labor demand is very strong. In labor market (b) the equilibrium wage is relatively low (W_b) because labor demand is weak. Clearly, the wage differential between occupations (a) and (b) results solely from differences in the magnitude of labor demand.

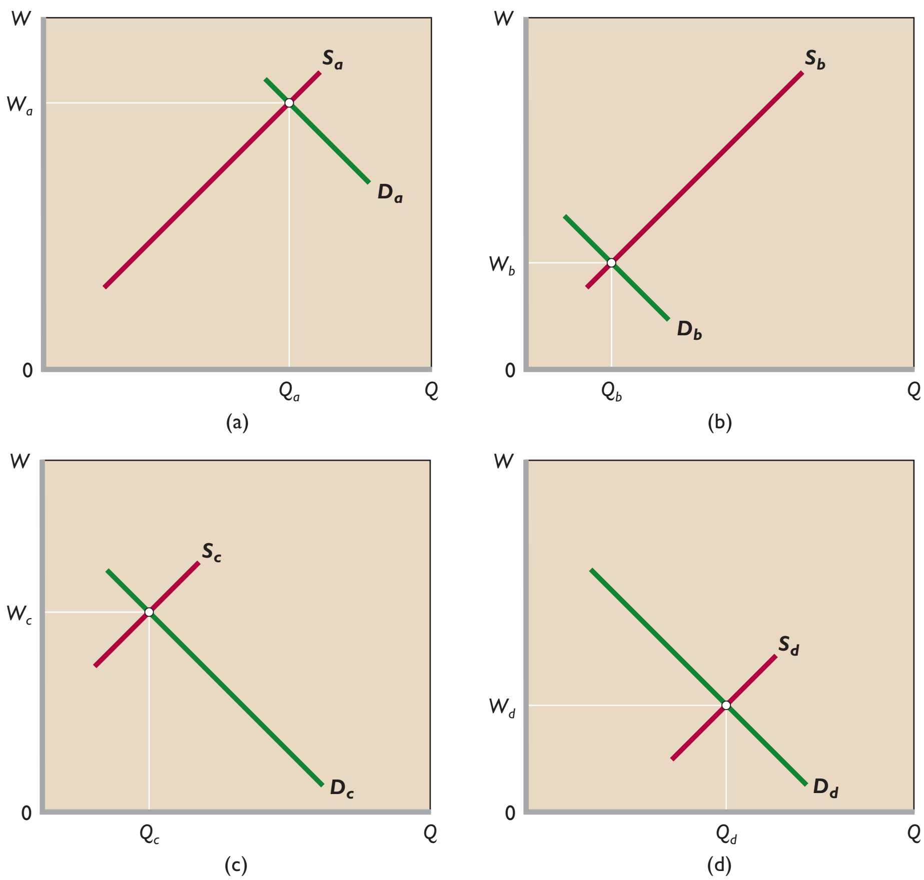


FIGURE 15.9 Labor demand, labor supply, and wage differentials. Wage differentials in labor markets are caused by varying supply and demand conditions. (a) and (b) Because the labor supply curves S_a and S_b are identical in the labor markets depicted in the two top graphs, differences in demand are the sole cause of the $W_a - W_b$ wage differential. (c) and (d) Because the labor demand curves D_c and D_d are identical in the bottom two graphs, the $W_c - W_d$ wage differential results solely from differences in labor supply.

Contrast that situation with Figure 15.9c and 15.9d, where the *labor demand curves* are identical. In labor market (c) the equilibrium wage is relatively high (W_c) because labor supply is low. In labor market (d) labor supply is highly abundant, so the equilibrium wage (W_d) is relatively low. The wage differential between (c) and (d) results solely from the differences in the magnitude of labor supply.

Although Figure 15.9 provides a good starting point for understanding wage differentials, we need to know *why* demand and supply conditions differ in various labor markets. There are several reasons.

Marginal Revenue Productivity

The strength of labor demand—how far rightward the labor demand curve is located—differs greatly among occupations due to differences in how much various occupational groups contribute to the revenue of their respective employers. This revenue contribution, in turn, depends on the workers' productivity and the strength of the demand for the products they are helping to produce. Where labor is highly productive and product demand is strong, labor demand also is strong and, other things equal, pay is high. Top professional athletes, for example, are highly productive at producing sports entertainment, for which millions of people are willing to pay billions of dollars over the course of a season. Because the **marginal revenue productivity** of these players is so high, they are in very high demand by sports teams. This high demand leads to their extremely high salaries (as in Figure 15.9a). In contrast, most workers generate much more modest revenue for their employers. This results in much lower demand for their labor and, consequently, much lower wages (as in Figure 15.9b).

Noncompeting Groups

On the supply side of the labor market, workers are not homogeneous; they differ in their mental and physical capacities and in their education and training. At any given time the labor force is made up of many **noncompeting groups** of workers, each representing several occupations for which the members of a particular group qualify. In some groups qualified workers are relatively few, whereas in others they are plentiful. And workers in one group do not qualify for the occupations of other groups.

Ability Only a few workers have the ability or physical attributes to be brain surgeons, concert violinists, top fashion models, research chemists, or professional athletes. Because the supply of these particular types of labor is very small in relation to labor demand, their wages are high (as in Figure 15.9c). The members of these and similar groups do not compete with one another or with other skilled or semiskilled workers. The violinist does not compete with the surgeon, nor does the surgeon compete with the violinist or the fashion model.

The concept of noncompeting groups can be applied to various subgroups and even to specific individuals in a particular group. Some especially skilled violinists can command higher salaries than colleagues who play the same instrument. A handful of top corporate executives earn 10 to 20 times as much as the average chief executive officer. In each of these cases, the supply of top talent is highly limited since less-talented colleagues are only imperfect substitutes.

Education and Training Another source of wage differentials is differing amounts of **human capital,** which is the personal stock of knowledge, know-how, and skills that enables a person to be productive and thus to earn income. Such stocks result from investments in human capital. Like expenditures on machinery and equipment, productivity-enhancing expenditures on education or training are investments. In both cases, people incur *present costs* with the intention that those expenditures will lead to a greater flow of *future earnings*.

ORIGIN OF THE IDEA

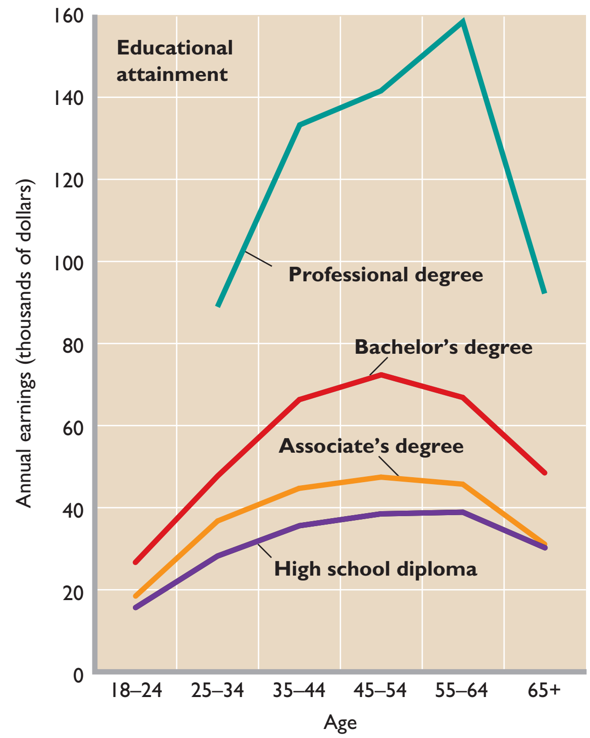
O15.2

Human capital

Figure 15.10 indicates that workers who have made greater investments in education achieve higher incomes during their careers. The reason is twofold: (1) There are fewer such workers, so their supply is limited relative to less-educated workers, and (2) more-educated workers tend to be more productive and thus in greater demand. Figure 15.10 also indicates that the earnings of better-educated workers rise more rapidly than those of poorly educated workers. The primary reason is that employers provide more on-the-job training to the better-educated workers, boosting their marginal revenue productivity and therefore their earnings.

Although education yields higher incomes, it carries substantial costs. A college education involves not only direct costs (tuition, fees, books) but indirect or opportunity costs (forgone earnings) as well. Does the higher pay received by better-educated workers compensate for these costs? The answer is yes. Rates of return are estimated to be 10 to 13 percent for investments in secondary education and 8 to 12 percent for investments in college education. One generally accepted estimate is that each year of schooling raises a worker's wage by about 8 percent.

FIGURE 15.10 Education levels and individual annual earnings. Annual income by age is higher for workers with more education than less. Investment in education yields a return in the form of earnings differences enjoyed over one's work life.



Source: U.S. Bureau of the Census. Data are for both sexes in 2011.

Compensating Differences

If the workers in a particular noncompeting group are equally capable of performing several different jobs, you might expect the wage rates to be identical for all these jobs. Not so. A group of high school graduates may be equally capable of becoming salesclerks or general construction workers. But these jobs pay different wages. In virtually all locales, construction laborers receive much higher wages than salesclerks. These wage differentials are called **compensating differences** because they must be paid to compensate for nonmonetary differences in various jobs.

The construction job involves dirty hands, a sore back, the hazard of accidents, and irregular employment, both seasonally and during recessions (the economywide economic slowdowns that periodically affect the economy). The retail sales job means clean clothing, pleasant air-conditioned surroundings, and little fear of injury or layoff. Other things equal, it is easy to see why workers would rather pick up a credit card than a shovel. So the amount of labor that is supplied to construction firms (as in Figure 15.9c) is smaller than that which is supplied to retail shops (as in Figure 15.9d). Construction firms must pay higher wages than retailers to compensate for the unattractive nonmonetary aspects of construction jobs.

Such compensating differences spring up throughout the economy. Other things equal, jobs having high risk of injury or death pay more than comparable, safer jobs. Jobs lacking employer-paid health insurance, pensions, and vacation time pay more than comparable jobs that provide these "fringe benefits." Jobs with more flexible hours pay less than jobs with rigid work-hour requirements. Jobs with greater risk of unemployment pay more than comparable jobs with little unemployment risk. Entry-level jobs in occupations that provide very poor prospects for pay advancement pay more than entry-level jobs that have clearly defined "job ladders."

CONSIDER THIS . . .

My Entire Life

Human capital is the accumulation of outcomes of prior investments in education, training, and other factors that increase productivity and earnings. It is the stock of knowledge, know-how, and skills that enables individuals to be productive and thus earn income. A valuable stock of human capital, together with a strong demand for one's services, can add up to a large capacity to earn income. For some people, high earnings have little to do with actual hours of work and much to do with their tremendous skill, which reflects their accumulated stock of human capital.

The point is demonstrated in the following story: It is said that a tourist once spotted the famous Spanish artist Pablo Picasso (1881–1973) in a Paris café. The tourist asked Picasso if he would do a sketch of his wife for pay. Picasso sketched the wife in a matter of minutes and said, "That will be 10,000 francs [roughly $2,000]." Hearing the high price, the tourist became irritated, saying, "But that took you only a few minutes."

"No," replied Picasso, "it took me my entire life!"

These and other compensating differences play an important role in allocating society's scarce labor resources. If very few workers want to be garbage collectors, then society must pay high wages to garbage collectors to get the garbage collected. If many more people want to be salesclerks, then society need not pay them as much as it pays garbage collectors to get those services performed.

Market Imperfections

Differences in marginal revenue productivity, amounts of human capital, and nonmonetary aspects of jobs explain most of the wage differentials in the economy. But some persistent differentials result from various market imperfections that impede workers from moving from lower-paying jobs to higher-paying jobs.

Lack of Job Information Workers may simply be unaware of job opportunities and wage rates in other geographic areas and in other jobs for which they qualify. Consequently, the flow of qualified labor from lower-paying to higher-paying jobs—and thus the adjustments in labor supply—may not be sufficient to equalize wages within occupations.

Geographic Immobility Workers take root geographically. Many are reluctant to move to new places. Doing so would involve leaving friends, relatives, and associates. It would mean forcing their children to change schools, having to sell their homes, and incurring the costs and inconveniences of adjusting to a new job and a new community. As Adam Smith noted over two centuries ago, "A [person] is of all sorts of luggage the most difficult to be transported." The reluctance or inability of workers to move enables geographic wage differentials within the same occupation to persist.

Unions and Government Restraints Wage differentials may be reinforced by artificial restrictions on mobility imposed by unions and government. We have noted that craft unions find it to their advantage to restrict membership. After all, if carpenters and bricklayers become too plentiful, the wages they can command will decline. Thus the low-paid nonunion carpenter of Brush, Colorado, may be willing to move to Chicago in the pursuit of higher wages. But her chances for succeeding are slim. She may be unable to get a union card, and no card means no job. Similarly, an optometrist or lawyer qualified to practice in one state may not meet the licensing requirements of other states, so his or her ability to move is limited. Other artificial barriers involve pension plans, health insurance benefits, and seniority rights that might be jeopardized by moving from one job to another.

Discrimination Despite legislation to the contrary, discrimination sometimes results in lower wages being paid to women and minority workers than to white males doing very similar or even identical work. Also, women and minorities may be crowded into certain low-paying occupations, driving down wages there and raising them elsewhere. If this *occupational segregation* keeps qualified women and minorities from taking higher-paying jobs, then differences in pay will persist. (We discuss discrimination in Chapter 21.)

All four considerations—differences in marginal revenue productivity, noncompeting groups, nonmonetary differences, and market imperfections—come into play in explaining actual wage differentials. For example, the differential between the wages of a physician and those of a construction worker can be explained on the basis of marginal revenue productivity and noncompeting groups. Physicians generate considerable revenue because of their high productivity and the strong willingness of consumers (via insurance) to pay for health care. Physicians also fall into a noncompeting group where, because of stringent training requirements, only relatively few persons qualify. So the supply of labor is small in relation to demand.

In construction work, where training requirements are much less significant, the supply of labor is great relative to demand. So wages are much lower for construction workers than for physicians. However, if not for the unpleasantness of the construction worker's job and the fact that his or her craft union observes restrictive membership policies, the differential would be even greater than it is.

Pay for Performance

LO15.8 Identify the types, benefits, and costs of "pay-for-performance" plans.

The models of wage determination we have described in this chapter assume that worker pay is always a standard amount for each hour's work, for example, $15 per hour. But pay schemes are often more complex than that both in composition and in purpose. For instance, many workers receive annual salaries rather than hourly pay. And workers receive differing proportions of fringe benefits (health insurance, life insurance, paid vacations, paid sick-leave days, pension contributions, and so on) as part of their pay. Finally, some pay plans are designed to elicit a desired level of performance from workers. This last aspect of pay plans requires further elaboration.

The Principal-Agent Problem

The **principal-agent problem** is usually associated with the possible differences in the interests of corporate stockholders (principals) and the executives (agents) they hire. But this problem extends to all paid employees. Firms hire workers because they are needed to help produce the goods and services the firms sell in their attempts to turn a profit. Workers are the firms' agents; they are hired to advance the interest (profit) of the firms. The principals are the firms; they hire agents to advance their goals. Firms and workers have one interest in common: They both want the firm to survive and thrive. That will ensure profit for the firm and continued employment and wages for the workers.

ORIGIN OF THE IDEA

O15.3

Principal-agent problem

But the interests of firms and workers are not identical. As a result, a principal-agent problem arises. Workers may seek to increase their utility by shirking on the job, that is, by providing less than the agreed-upon effort or by taking unauthorized breaks. They may improve their well-being by increasing their leisure during paid work hours, without forfeiting income. The night security guard in a warehouse may leave work early or spend time reading a novel rather than making the assigned rounds. A salaried manager may spend time away from the office visiting with friends rather than attending to company business.

Firms (principals) have a profit incentive to reduce or eliminate shirking. One option is to monitor workers, but monitoring is difficult and costly. Hiring another worker to supervise or monitor the security guard might double the cost of maintaining a secure warehouse. Another way of resolving a principal-agent problem is through some sort of **incentive pay plan** that ties worker compensation more closely to worker output or performance. Such incentive pay schemes include piece rates; commissions and royalties; bonuses, stock options, and profit sharing; and efficiency wages.

Piece Rates Piece rates consist of compensation paid according to the number of units of output a worker produces. If a principal pays fruit pickers by the bushel or typists by the page, it need not be concerned with shirking or with monitoring costs.

Commissions or Royalties Unlike piece rates, commissions and royalties tie compensation to the value of sales. Employees who sell products or services—including real estate agents, insurance agents, stockbrokers, and retail salespersons—commonly receive *commissions* that are computed as a percentage of the monetary value of their sales. Recording artists and authors are paid *royalties*, computed as a certain percentage of sales revenues from their works. Such types of compensation link the financial interests of the salespeople, artists, and authors to the profit interest of the firms.

Bonuses, Stock Options, and Profit Sharing *Bonuses* are payments in addition to one's annual salary that are based on some factor such as the performance of the individual worker, or of a group of workers, or of the firm itself. A professional baseball player may receive a bonus based on a high batting average, the number of home runs hit, or the number of runs batted in. A business manager may receive a bonus based on the profitability of her or his unit. *Stock options* allow workers to buy shares of their employer's stock at a fixed, lower price when the stock price rises. Such options are part of the compensation packages of top corporate officials, as well as many workers in relatively high-technology firms. *Profit-sharing plans* allocate a percentage of a firm's profit to its employees.

Efficiency Wages The rationale behind *efficiency wages* is that employers will enjoy greater effort from their workers by paying them above-equilibrium wage rates. Glance back at Figure 15.3, which shows a competitive labor market in which the equilibrium wage rate is \$10. What if an employer decides to pay an above-equilibrium wage of \$12 per hour? Rather than putting the firm at a cost disadvantage compared with rival firms paying only \$10, the higher wage might improve worker effort and productivity so that unit labor costs actually fall. For example, if each worker produces 10 units of output per hour at the \$12 wage rate compared with only 6 units at the \$10 wage rate, unit labor costs for the high-wage firm will be only \$1.20 (= \$12/10) compared to \$1.67 (= \$10/6) for firms paying the equilibrium wage.

An above-equilibrium wage may enhance worker efficiency in several ways. It enables the firm to attract higher-quality workers. It lifts worker morale. And it lowers turnover, resulting in a more experienced workforce, greater worker productivity, and lower recruitment and training costs. Because the opportunity cost of losing a higher-wage job is greater, workers are more likely to put forth their best efforts with less supervision and monitoring. In fact, efficiency wage payments have proved effective for many employers.

ORIGIN OF THE IDEA

O15.4

Efficiency wages

Addenda: Negative Side Effects of Pay for Performance

Although pay for performance may help overcome the principal-agent problem and enhance worker productivity, such plans may have negative side effects and require careful design. Here are a few examples:

- The rapid production pace that piece rates encourage may result in poor product quality and may compromise the safety of workers. Such outcomes can be costly to the firm over the long run.
- Commissions may cause some salespeople to engage in questionable or even fraudulent sales practices, such as making exaggerated claims about products or recommending unneeded repairs. Such practices may lead to private lawsuits or government legal action.
- Bonuses based on personal performance may disrupt the close cooperation needed for maximum team production. A professional basketball player who receives a bonus for points scored may be reluctant to pass the ball to teammates.
- Since profit sharing is usually tied to the performance of the entire firm, less energetic workers can "free ride" by obtaining their profit share on the basis of the hard work by others.
- Stock options may prompt some unscrupulous executives to manipulate the cost and revenue streams of their firms to create a false appearance of rapidly rising profit. When the firm's stock value rises, the executives exercise their stock options at inflated share prices and reap a personal fortune.
- There may be a downside to the reduced turnover resulting from above-market wages: Firms that pay efficiency wages have fewer opportunities to hire new workers and suffer the loss of the creative energy that they often bring to the workplace.

QUICK REVIEW 15.3

- Proponents of the minimum wage argue that it is needed to assist the working poor and to counter monopsony where it might exist; critics say that it is poorly targeted to reduce poverty and that it reduces employment.
- Wage differentials are attributable in general to the forces of supply and demand, influenced by differences in workers' marginal revenue productivity, education, and skills and by nonmonetary differences in jobs. But several labor market imperfections also play a role.
- As it applies to labor, the principal-agent problem is one of workers pursuing their own interests to the detriment of the employer's profit objective.
- Pay-for-performance plans (piece rates, commissions, royalties, bonuses, stock options, profit sharing, and efficiency wages) are designed to improve worker productivity by overcoming the principal-agent problem.

LAST WORD

Are Chief Executive Officers (CEOs) Overpaid?

The Multimillion-Dollar Pay of Major Corporate CEOs Has Drawn Considerable Criticism.

Top executives of U.S. corporations typically receive total annual pay (salary, bonuses, and stock options) in the millions of dollars. As shown in Table 1, each of the five highest-paid U.S. executives earned more than $50 million in 2011.

CEO pay in the United States is not only exceptionally high relative to the average pay of U.S. managers and workers but also high compared to the CEO pay in other industrial countries. For example, in 2005 the CEO pay at firms with about $500 million in annual sales averaged $2.2 million in the United States, compared to $1.2 million in France and Germany and less than $600,000 in South Korea and Japan.*

Is high CEO pay simply the outcome of labor supply and labor demand, as is the pay for star athletes and entertainers? Does it reflect marginal revenue productivity—that is, the contributions by CEOs to their company's output and revenue?

Observers who answer affirmatively point out that decisions made by the CEOs of large corporations affect the productivity of every employee in the organization. Good decisions enhance productivity throughout the organization and increase revenue; bad decisions reduce productivity and revenue. Only executives who have consistently made good business decisions attain the top positions in large corporations. Because the supply of these people is highly limited and their marginal revenue productivity is enormous, they command huge salaries and performance bonuses.

Also, some economists note that CEO pay in the United States may be like the prizes professional golfers and tennis players receive for winning tournaments. These high prizes are designed to promote the productivity of all those who aspire to achieve them. In corporations the top prizes go to the winners of the "contests" among managers to attain, at least eventually, the CEO positions. Thus high CEO pay does not derive solely from the CEO's direct productivity. Instead, it may exist because the high pay creates incentives that raise the productivity of scores of other corporate executives who seek to achieve the top position. In this view, high CEO pay remains grounded on high productivity.

Critics of existing CEO pay acknowledge that CEOs deserve substantially higher salaries than ordinary workers or typical managers, but they question pay packages that run into the millions of dollars. They reject the "tournament pay" idea on the grounds that corporations require cooperative team effort by managers and executives, not the type of high-stakes competition promoted by "winner-take-most" compensation. They believe that corporations, although owned by their shareholders, are controlled by corporate boards and professional executives. Because many board members are present or past CEOs of other corporations, they often exaggerate CEO importance and, consequently, overpay their own CEOs. These overpayments are at the expense of the firm's stockholders.

In summary, defenders of CEO pay say that high pay is justified by the direct or indirect marginal-revenue contribution of CEOs. Like it or not, CEO pay is market-determined pay. In contrast, critics say that multimillion-dollar CEO pay bears little relationship to marginal revenue productivity and is unfair to ordinary stockholders. It is clear from our discussion that this issue remains unsettled.

TABLE 1 The Five Highest-Paid U.S. CEOs, 2011

Name	Company	Total Pay, Millions
John H. Hammergren	McKesson	$131
Ralph Lauren	Ralph Lauren	67
Michael D. Fascitelli	Vornado Realty	64
Richard G. Kinder	Kinder Morgan	61
David M. Cote	Honeywell	56

Source: Forbes, **www.forbes.com**. Reprinted by permission of Forbes Media LLC © 2012.

**Worldwide Total Remuneration, 2005–2006* (New York: Towers Perrin, Jan. 11, 2006, p. 20).

SUMMARY

LO15.1 Explain why labor productivity and real hourly compensation track so closely over time.

The term "labor" encompasses all people who work for pay. The wage rate is the price paid per unit of time for labor. Labor earnings comprise total pay and are found by multiplying the number of hours worked by the hourly wage rate. The nominal wage rate is the amount of money received per unit of time; the real wage rate is the purchasing power of the nominal wage.

The long-run growth of real hourly compensation—the average real wage—roughly matches that of productivity, with both increasing over the long run.

Global comparisons suggest that real wages in the United States are relatively high, but not the highest, internationally. High real wages in the advanced industrial countries stem largely from high labor productivity.

LO15.2 Show how wage rates and employment levels are determined in competitive labor markets.

Specific wage rates depend on the structure of the particular labor market. In a competitive labor market the equilibrium wage rate and level of employment are determined at the intersection of the labor supply curve and labor demand curve. For the individual firm, the market wage rate establishes a horizontal labor supply curve, meaning that the wage rate equals the firm's constant marginal resource cost. The firm hires workers to the point where its MRP equals its MRC.

LO15.3 Demonstrate how monopsony (a market with a single employer) can reduce wages below competitive levels.

Under monopsony the marginal resource cost curve lies above the resource supply curve because the monopsonist must bid up the wage rate to hire extra workers and must pay that higher wage rate to all workers. The monopsonist hires fewer workers than are hired under competitive conditions, pays less-than-competitive wage rates (has lower labor costs), and thus obtains greater profit.

LO15.4 Discuss how unions increase wage rates by pursuing the demand-enhancement model, the craft union model, or the industrial union model.

A union may raise competitive wage rates by (*a*) increasing the derived demand for labor, (*b*) restricting the supply of labor through exclusive unionism, or (*c*) directly enforcing an above-equilibrium wage rate through inclusive unionism.

LO15.5 Explain why wages and employment are determined by collective bargaining in a situation of bilateral monopoly.

In many industries the labor market takes the form of bilateral monopoly, in which a strong union "sells" labor to a monopsonistic employer. The wage-rate outcome of this labor market model depends on union and employer bargaining power.

LO15.6 Discuss how minimum wage laws affect labor markets.

On average, unionized workers realize wage rates 15 percent higher than those of comparable nonunion workers.

Economists disagree about the desirability of the minimum wage as an antipoverty mechanism. While it causes unemployment for some low-income workers, it raises the incomes of those who retain their jobs.

LO15.7 List the major causes of wage differentials.

Wage differentials are largely explainable in terms of (*a*) marginal revenue productivity of various groups of workers; (*b*) noncompeting groups arising from differences in the capacities and education of different groups of workers; (*c*) compensating wage differences, that is, wage differences that must be paid to offset nonmonetary differences in jobs; and (*d*) market imperfections in the form of lack of job information, geographic immobility, union and government restraints, and discrimination.

LO15.8 Identify the types, benefits, and costs of "pay-for-performance" plans.

As it a applies to labor, the principal-agent problem arises when workers provide less-than-expected effort. Firms may combat this by monitoring workers or by creating incentive pay schemes that link worker compensation to performance.

TERMS AND CONCEPTS

wage rate
nominal wage
real wage
purely competitive labor market
monopsony
exclusive unionism
occupational licensing
inclusive unionism
bilateral monopoly
minimum wage
wage differentials
marginal revenue productivity
noncompeting groups
human capital
compensating differences
principal-agent problem
incentive pay plan

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Explain why the general level of wages is high in the United States and other industrially advanced countries. What is the single most important factor underlying the long-run increase in average real-wage rates in the United States? **LO15.1**
2. Why is a firm in a purely competitive labor market a wage taker? What would happen if it decided to pay less than the going market wage rate? **LO15.2**
3. Describe wage determination in a labor market in which workers are unorganized and many firms actively compete for the services of labor. Show this situation graphically, using W_1 to indicate the equilibrium wage rate and Q_1 to show the number of workers hired by the firms as a group. Show the labor supply curve of the individual firm, and compare it with that of the total market. Why the differences? In the diagram representing the firm, identify total revenue, total wage cost, and revenue available for the payment of nonlabor resources. **LO15.2**
4. Suppose the formerly competing firms in the previous question form an employers' association that hires labor as a monopsonist would. Describe verbally the effect on wage rates and employment. Adjust the graph you drew for review question 1, showing the monopsonistic wage rate and employment level as W_2 and Q_2, respectively. Using this monopsony model, explain why hospital administrators sometimes complain about a "shortage" of nurses. How might such a shortage be corrected? **LO15.3**
5. Assume a monopsonistic employer is paying a wage rate of W_m and hiring Q_m workers, as indicated in Figure 15.8. Now suppose an industrial union is formed that forces the employer to accept a wage rate of W_c. Explain verbally and graphically why in this instance the higher wage rate will be accompanied by an increase in the number of workers hired. **LO15.5**
6. Have you ever worked for the minimum wage? If so, for how long? Would you favor increasing the minimum wage by a dollar? By two dollars? By five dollars? Explain your reasoning. **LO15.6**
7. "Many of the lowest-paid people in society—for example, short-order cooks—also have relatively poor working conditions. Hence, the notion of compensating wage differentials is disproved." Do you agree? Explain. **LO15.7**
8. What is meant by investment in human capital? Use this concept to explain (*a*) wage differentials and (*b*) the long-run rise of real-wage rates in the United States. **LO15.7**
9. What is the principal-agent problem? Have you ever worked in a setting where this problem has arisen? If so, do you think increased monitoring would have eliminated the problem? Why don't firms simply hire more supervisors to eliminate shirking? **LO15.8**
10. **LAST WORD** Do you think exceptionally high pay to CEOs is economically justified? Why or why not?

REVIEW QUESTIONS

1. Brenda owns a construction company that employs bricklayers and other skilled tradesmen. Her firm's MRP for bricklayers is \$22.25 per hour for each of the first seven bricklayers, \$18.50 for an eighth brick layer, and \$17.75 for a ninth bricklayer. Given that she is a price taker when hiring bricklayers, how many bricklayers will she hire if the market equilibrium wage for bricklayers is \$18.00 per hour? **LO15.2**
   - **a.** Zero.
   - **b.** Seven.
   - **c.** Eight.
   - **d.** Nine.
   - **e.** More information is required to answer this question.
2. Because a perfectly competitive employer's MRC curve is ______________, it will hire ______________ workers than would a monoposony employer with the same MRP curve. **LO15.3**
   - **a.** Upsloping; more.
   - **b.** Upsloping; fewer.
   - **c.** Flat; more.
   - **d.** Flat; fewer.
   - **e.** Downsloping; more.
   - **f.** Downsloping; fewer.
3. True or false. When a labor market consists of a single monopsony buyer of labor interacting with a single monopoly seller of labor (such as a trade union), the resulting quantity of labor that is hired will always be inefficiently low. **LO15.5**
4. The market equilibrium wage is currently \$12 per hour among hairdressers. At that wage, 17,323 hairdressers are currently employed in the state. The state legislature then sets a minimum wage of \$11.50 per hour for hairdressers. If there are no changes to either the demand or supply for hairdressers when that minimum wage is imposed, the number of hairdressers employed in the state will be: **LO15.6**
   - **a.** Fewer than 17,323.
   - **b.** Still 17,323.
   - **c.** More than 17,323.
   - **d.** This is a bilateral monopsony so you can't tell.
5. On average, 50-year-old workers are paid several times more than workers in their teens and twenties. Which of the following options is the most likely explanation for that huge difference in average earnings? **LO15.7**
   - **a.** Older workers have more human capital and higher MRPs.

b. Employers engage in widespread discrimination against younger workers.
c. Young people lack information about the existence of the high-paying jobs occupied by older workers.
d. Older workers receive compensating differences because they do jobs that are more risky than the jobs done by younger workers.

6. Manny owns a local fast-food franchise. Angel runs it for him. So in this situation, Manny is the ____________ and Angel is the ____________. **LO15.8**
a. Free rider; entrepreneur.
b. Agent; principal.
c. Principal; agent.
d. Producer; consumer.

7. A principal is worried that her agent may not do what she wants. As a solution, she should consider: **LO15.8**
a. Commissions.
b. Bonuses.
c. Profit sharing.
d. All of the above.

PROBLEMS

1. Workers are compensated by firms with "benefits" in addition to wages and salaries. The most prominent benefit offered by many firms is health insurance. Suppose that in 2000, workers at one steel plant were paid $20 per hour and in addition received health benefits at the rate of $4 per hour. Also suppose that by 2010 workers at that plant were paid $21 per hour but received $9 in health insurance benefits. **LO15.1**
a. By what percentage did total compensation (wages plus benefits) change at this plant from 2000 to 2010? What was the approximate average annual percentage change in total compensation?
b. By what percentage did wages change at this plant from 2000 to 2010? What was the approximate average annual percentage change in wages?
c. If workers value a dollar of health benefits as much as they value a dollar of wages, by what total percentage will they feel that their incomes have risen over this time period? What if they only consider wages when calculating their incomes?
d. Is it possible for workers to feel as though their wages are stagnating even if total compensation is rising?

2. Complete the following labor supply table for a firm hiring labor competitively: **LO15.2**

Units of Labor	Wage Rate	Total Labor Cost	Marginal Resource (Labor) Cost
0	$14	$ ______	
			$ ______
1	14	______	
			______
2	14	______	
			______
3	14	______	
			______
4	14	______	
			______
5	14	______	
			______
6	14	______	

a. Show graphically the labor supply and marginal resource (labor) cost curves for this firm. Are the curves the same or different? If they are different, which one is higher?
b. Plot the labor demand data of review question 2 in Chapter 14 on the graph used in part *a* above. What are the equilibrium wage rate and level of employment?

3. Assume a firm is a monopsonist that can hire its first worker for $6 but must increase the wage rate by $3 to attract each successive worker (so that the second worker must be paid $9, the third $12, and so on). **LO15.3**
a. Draw the firm's labor supply and marginal resource cost curves. Are the curves the same or different? If they are different, which one is higher?
b. On the same graph, plot the labor demand data of review question 2 in Chapter 14. What are the equilibrium wage rate and level of employment?
c. Compare these answers with those you found in problem 2. By how much does the monopsonist reduce wages below the competitive wage? By how much does the monopsonist reduce employment below the competitive level?

4. Suppose that low-skilled workers employed in clearing woodland can each clear one acre per month if each is equipped with a shovel, a machete, and a chainsaw. Clearing one acre brings in $1,000 in revenue. Each worker's equipment costs the worker's employer $150 per month to rent and each worker toils 40 hours per week for four weeks each month. **LO15.6**
a. What is the marginal revenue product of hiring one low-skilled worker to clear woodland for one month?
b. How much revenue per hour does each worker bring in?
c. If the minimum wage were $6.20, would the revenue per hour in part *b* exceed the minimum wage? If so, by how much per hour?
d. Now consider the employer's total costs. These include the equipment costs as well as a normal profit of $50 per acre. If the firm pays workers the minimum wage of $6.20 per hour, what will the firm's economic profit or loss be per acre?
e. At what value would the minimum wage have to be set so that the firm would make zero economic profit from employing an additional low-skilled worker to clear woodland?

5. Suppose that a car dealership wishes to see if efficiency wages will help improve its salespeople's productivity. Currently, each salesperson sells an average of one car per

day while being paid \$20 per hour for an eight-hour day. **LO15.8**

a. What is the current labor cost per car sold?

b. Suppose that when the dealer raises the price of labor to \$30 per hour the average number of cars sold by a salesperson increases to two per day. What is now the labor cost per car sold? By how much is it higher or lower than it was before? Has the efficiency of labor expenditures by the firm (cars sold per dollar of wages paid to salespeople) increased or decreased?

c. Suppose that if the wage is raised a second time to \$40 per hour the number of cars sold rises to an average of 2.5 per day. What is now the labor cost per car sold?

d. If the firm's goal is to maximize the efficiency of its labor expenditures, which of the three hourly salary rates should it use: \$20 per hour, \$30 per hour, or \$40 per hour?

e. By contrast, which salary maximizes the productivity of the car dealer's workers (cars sold per worker per day)?

Labor Unions and Their Impacts

LO15.9 Relate who belongs to U.S. unions, the basics of collective bargaining, and the economic effects of unions.

We have noted that unions can increase wage rates by augmenting the demand for labor (Figure 15.5) or by restricting or controlling the supply of labor (Figures 15.6 and 15.7). The purpose of this appendix is to provide some additional information about American unions, collective bargaining, and union impacts.

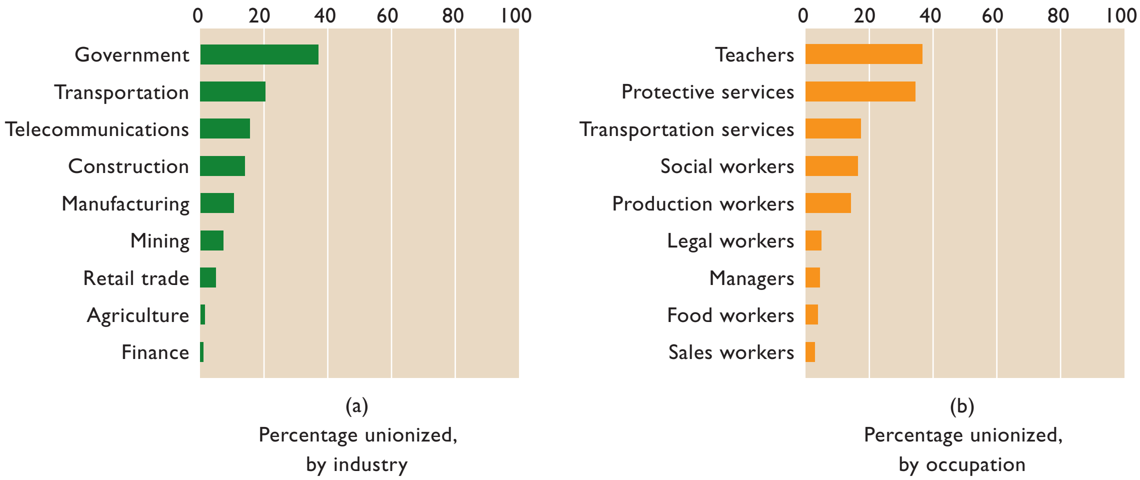
Union Membership

In 2011, about 14.8 million U.S. workers—11.8 percent of employed wage and salary workers—belonged to unions. Some 8.4 million of these U.S. union members belonged to one of many unions that are loosely and voluntarily affiliated with the **American Federation of Labor and the Congress of Industrial Organizations (AFL-CIO).** Examples of AFL-CIO unions are the United Autoworkers, Communications Workers, and United Steelworkers. Another 5.5 million union members belonged to one of the seven unions, including the Service Workers and Teamsters, loosely federated as **Change to Win.** The remaining union members belonged to other **independent unions** that were not affiliated with either federation.

The likelihood that any particular worker will be a union member depends mainly on the industry in which the worker is employed and his or her occupation. As shown in Figure 1a, the **unionization rate**—the percentage of workers unionized—is high in government, transportation, telecommunications, construction, and manufacturing. The unionization rate is very low in finance, agriculture, and retail trade. Figure 1b shows that unionism also varies greatly by occupation. Teachers, protective service workers, transportation workers, production workers, and social workers have high unionization rates; sales workers, food workers, and managers have very low rates.

Because disproportionately more men than women work in the industries and occupations with high unionization rates, men are more likely to be union members than women. Specifically, 12 percent of male wage and salary workers belong to unions compared with 11 percent of women. For the same reason, African Americans have higher unionization rates than whites: 14 percent compared with 12 percent. The unionization rate for Asians is 10 percent; Hispanics, 10 percent. Unionism in the United States is largely an urban phenomenon. Six heavily urbanized, heavily industrialized states (New York, California, Pennsylvania, Illinois, Ohio, and Michigan) account for approximately half of all union members.

FIGURE 1 Union membership as a percentage of employed wage and salary workers, selected industries and occupations, 2011. In percentage terms, union membership varies greatly by (a) industry and (b) occupation.



Source: Bureau of Labor Statistics, **www.bls.gov**.

The Decline of Unionism

Since the mid-1950s, union membership has not kept pace with the growth of the labor force. While 25 percent of employed wage and salary workers belonged to unions in the mid-1950s, today only 11.8 percent are union members. Over recent years, even the absolute number of union members has declined significantly. More than 22 million workers were unionized in 1980 but only 14.8 million in 2011.

Some of the major reasons for the decline of U.S. unionism involve structural changes in the economy. Employment has shifted away from manufactured goods (where unions have been stronger) and toward services (where unions have been weaker). Consumer demand has shifted toward foreign manufactured goods and away from goods produced by union labor in the United States. Industry has shifted from the northeast and midwest, where unionism is "a way of life," to "hard-to-organize" areas of the south and southwest. These and other factors have reduced the growth of union membership.

Also, management has greatly intensified its opposition to unions and has increasingly engaged in aggressive collective bargaining, including the use of strikebreakers. Within unionized firms, employers have substituted machinery for workers, subcontracted work to nonunion suppliers, and shifted the production of components to low-wage nations. Nonunion firms have greatly improved their wage, fringe benefits, and working conditions. That has reduced the demand for unionism.

Collective Bargaining

Despite the overall decline of unionism, **collective bargaining** (the negotiation of labor contracts) remains an important feature of labor-management relations in several U.S. industries. The goal of collective bargaining is to establish a "work agreement" between the firm and the union.

Collective bargaining agreements (contracts) assume many forms, but typically cover several topics.

Union Status

Union status is the degree of security afforded a union by the work agreement. The strongest form of union security is a **closed shop,** in which a worker must be (or must become) a member of the union before being hired. Under federal labor law, such shops are illegal in industries other than transportation and construction.

In contrast, a **union shop** permits the employer to hire nonunion workers but provides that these workers must join the union within a specified period, say 30 days, or relinquish their jobs. An **agency shop** allows nonunion workers but requires nonunion workers to either pay union dues or donate an equivalent amount to charity. Union and agency shops are legal, except in the 22 states that expressly prohibit them through so-called **right-to-work laws.**

In an **open shop,** an employer may hire either union or nonunion workers. Those who are nonunion are not obligated to join the union or to pay union dues; they may continue on their jobs indefinitely as nonunion workers. Nevertheless, the wages, hours, and working conditions set forth in the work agreement apply to the nonunion workers as well as to the union workers.

Managerial Prerogatives

Most work agreements contain clauses outlining certain decisions that are reserved solely for management. These managerial prerogatives usually cover such matters as the size and location of plants, the products to be manufactured, and the types of equipment and materials to be used in production and in production scheduling.

Wages and Hours

The focal point of almost all bargaining agreements is wages (including fringe benefits) and hours. Both labor and management press for the advantage in wage bargaining. The arguments that unions use most frequently in demanding wage boosts are (1) "what others are getting"; (2) the employer's ability to pay, based on its profitability; (3) increases in the cost of living; and (4) increases in labor productivity.

Hours of work, voluntary versus mandatory overtime, holiday and vacation provisions, profit sharing, health plans, and pension benefits are other contract issues that must be addressed in the bargaining process.

Seniority and Job Protection

The uncertainty of employment in a market economy, along with the fear of antiunion discrimination on the part of employers, has made workers and their unions "job-conscious." The explicit and detailed provisions covering job opportunities that most agreements contain reflect this concern. Unions stress length of service, or *seniority*, as the basis for worker promotion and for layoff and recall. They want the worker with the longest continuous service to have the first chance at relevant promotions, to be the last one laid off, and to be the first one recalled from layoff.

In recent years, unions have become increasingly sensitive to losing jobs to nonunion subcontractors and to overseas workers. Unions sometimes seek limits on the firm's ability to subcontract out work or to relocate production facilities overseas.

Grievance Procedures

Even the most detailed and comprehensive work agreement cannot spell out all the specific issues and problems that might occur during its life. For example, suppose a particular worker gets reassigned to a less pleasant job. Was this reassignment for legitimate business reasons or, as the person suspects, because of a personality conflict with a particular manager? Labor contracts contain grievance procedures to resolve such matters.

The Bargaining Process

The date for the beginning of collective bargaining on a new contract is usually specified in the existing contract and is typically 60 days before the current one expires.

The union normally takes the initiative, presenting its demands in the form of specific wage, fringe-benefit, and other adjustments to the present union-management contract. The firm counters with an offer relating to these and other contract provisions. It is not unusual for the original union demand and the first offer by the firm to be far apart, not only because of the parties' conflicting goals but also because starting far apart leaves plenty of room for compromise and counteroffers during negotiations.

Hanging over the negotiations is the contract deadline, which occurs the moment the present contract expires. At that time there is a possibility of a **strike**—a work stoppage by the union—if it thinks its demands are not being satisfactorily met. But there is also the possibility that at the deadline the firm may engage in a **lockout,** in which it forbids the workers to return to work until a new contract is signed. In this setting of uncertainty prior to the deadline, both parties feel pressure to find mutually acceptable terms.

Although bluster and bickering often occur in collective bargaining, labor and management display a remarkable capacity for compromise and agreement. Typically they reach a compromise that is written into a new contract. Nevertheless, strikes and lockouts occasionally do occur. When they happen, workers lose income and firms lose profit. To stem their losses, both parties usually look for and eventually find ways to settle the labor dispute and get the workers back to work.

Bargaining, strikes, and lockouts occur within a framework of federal labor law, specifically the **National Labor Relations Act (NLRA).** This act was first passed as the Wagner Act of 1935 and later amended by the Taft-Hartley Act of 1947 and the Landrum-Griffin Act of 1959. The act sets forth the *dos and don'ts* of union and management labor practices. For example, while union members can picket in front of a firm's business, they cannot block access to the business by customers, coworkers, or strikebreakers hired by the firm. Another example: Firms cannot refuse to meet and talk with the union's designated representatives.

Either unions or management can file charges of unfair labor practices under the labor law. The **National Labor Relations Board (NLRB)** has the authority to investigate such charges and to issue cease-and-desist orders in the event of a violation. The board also conducts worker elections to decide which specific union, if any, a group of workers might want to have represent them in collective bargaining.

Economic Effects of Unions

The most straightforward effect of unions is an increase in the wage rates for their members. The consensus estimate is that the overall union wage premium (wage advantage) averages about 15 percent. The effects of unions on output and efficiency, however, are slightly more complicated.

Featherbedding and Work Rules

Some unions diminish output and efficiency by engaging in "make-work" or "featherbedding" practices and resisting the introduction of output-increasing machinery and equipment. These productivity-reducing practices often arise in periods of technological change. For example, in 2002 the ILWU (dockworkers' union) obtained a contract provision guaranteeing 40-hour-per-week jobs for all current ILWU clerical personnel for as long as they wish to continue working at their current jobs at West Coast ports. However, many of those workers will not be needed because the ports are rapidly moving toward labor-saving computerized systems for tracking cargo. Thus, many of the current clerical personnel will be paid for doing little or nothing. This will be very inefficient.

More generally, unions may reduce efficiency by establishing work rules and practices that impede putting the most productive workers in particular jobs. Under seniority rules, for example, workers may be promoted for their employment tenure rather than for their ability to perform the available job with the greatest efficiency. Also, unions might restrict the kinds of tasks workers may perform. Contract provisions may prohibit sheet-metal workers or bricklayers from doing the simple carpentry work often associated with their jobs. Observance of such rules means, in this instance, that firms must hire unneeded and under-used carpenters.

Finally, critics of unions contend that union contracts often chip away at managerial prerogatives to establish work schedules, determine production targets, introduce new technology, and make other decisions contributing to productive efficiency.

Output Losses from Strikes

A second way unions can impair efficiency and output is through strikes. If union and management reach an impasse during their contract negotiations, a strike may result and the firm's production may cease for the strike's duration. If so, the firm will forgo sales and profit; workers will sacrifice income; and the economy might lose output. U.S. strike activity, however, has dwindled in the past few decades. In 2011, there were 19 major work stoppages—strikes or lockouts involving 1,000 or more employees.

About 113,000 workers were idled by the 19 work stoppages in 2011, with the average length of stoppages being 18 days. It is estimated that the amount of work time lost to the stoppages was less than 0.005 percent of the total work time provided by employees that year.

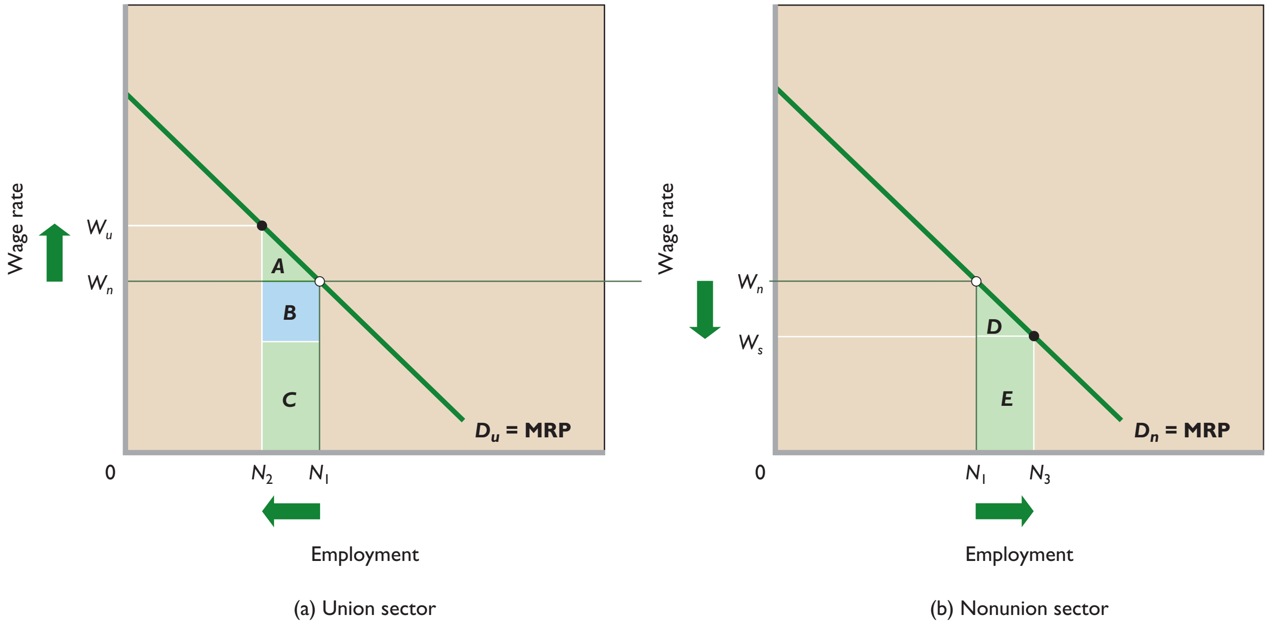
But the amount of work time lost is an imprecise indicator of the potential economic costs of strikes. These costs may be greater than indicated if strikes disrupt production in nonstruck firms that either supply inputs to struck firms or buy products from them. Example: An extended strike in the auto industry might reduce output and cause layoffs in firms producing, say, glass, tires, paints, and fabrics used in producing cars. It also may reduce sales and cause layoffs in auto dealerships.

On the other hand, the costs of strikes may be less than is implied by the work time lost by strikers if nonstruck firms increase their output to offset the loss of production by struck firms. While the output of General Motors declines when its workers strike, auto buyers may shift their demand to Ford, Toyota, or Honda, which will respond by increasing their employment and output. Therefore, although GM and its employees are hurt by a strike, society as a whole may experience little or no decline in employment, real output, and income.

Efficiency Losses from Labor Misallocation

A third and more subtle way that unions might reduce efficiency and output is through the union wage advantage itself. Figure 2 splits the economy into two sectors, showing identical labor demand curves for the unionized sector and the nonunionized sector. If all markets are competitive and no union is initially present in either sector, the wage rate in both parts of the economy will be W_n and N_1 workers will be employed in each sector.

FIGURE 2 The effects of the union wage advantage on the allocation of labor. (a) The higher wage rate W_u obtained by unions in the union sector displaces $N_1 - N_2$ workers and reduces output by area $A + B + C$. (b) The displaced workers from the union sector are reemployed in the nonunion sector, where employment increases from N_1 to N_3 and the wage rate declines from W_n to W_s. Output increases by area $D + E$ in the nonunion sector, but that increase is less than the loss of output $A + B + C$ in the union sector. The net loss of output, shown by area B in blue, means that the union wage advantage has caused a misallocation of resources and a decline in economic efficiency.



Now suppose workers form a union in one sector and succeed in increasing the wage rate from W_n to W_u. As a consequence, $N_1 - N_2$ workers lose their jobs in the union sector. Assume that they all move to the nonunion sector, where they are employed. This increase in labor supply (not shown) in the nonunion sector increases the quantity of labor supplied there from N_1 to N_3, reducing the wage rate from W_n to W_s.

Recall that the labor demand curves reflect the marginal revenue products (MRPs) of workers or, in other words, the contribution that each additional worker makes to domestic output. This means that area A + B + C in the union sector represents the sum of the MRPs—the total contributions to domestic output—of the workers displaced by the wage increase achieved by the union. The reemployment of these workers in the nonunion sector produces an increase in domestic output shown by area D + E. Because area A + B + C exceeds area D + E, a net loss of domestic output is the result.

More precisely, because A = D and C = E, the efficiency loss attributable to the union wage advantage is represented by area B. Because the same amount of employed labor is now producing a smaller output, labor is being misallocated and used inefficiently. After the shift of N_1N_2 workers to the nonunion sector has occurred, workers in both sectors will be paid wage rates according to their MRPs. But the workers who shifted sectors will be working at a lower MRP than before. An economy always obtains a larger domestic output when labor is reallocated from a low MRP use to a high-MRP use. But here the opposite has occurred. And assuming the union can maintain the W_u wage rate in its sector, a reallocation of labor from the nonunion sector to the union sector will never occur.

Attempts to estimate the efficiency loss associated with union wage gains, however, suggest that it is very small: perhaps 0.2 to 0.4 percent (or one-fifth of 1 percent to two-fifths of 1 percent) of U.S. GDP. In 2011 this cost would have amounted to about $32 billion to $63 billion.

Offsetting Factors

Some long-run consequences of unionization may enhance productivity and reduce the efficiency loss from unions. One such impact is lower worker turnover within unionized firms. Compared with the rates at nonunion firms, the quit rates (resignation rates) for union workers are 31 to 65 percent lower, depending on the industry.

The union wage premium may reduce worker turnover by increasing the desirability of the union job relative to alternative employment. In economic terms, the higher opportunity cost of quitting reduces the frequency of quitting. Unions also may reduce turnover by using collective communication—the **voice mechanism**—to correct job dissatisfactions that otherwise would be "resolved" by workers quitting and taking other jobs—the **exit mechanism.** It might be risky for individual workers to express their dissatisfaction to employers because employers might retaliate by firing them as "troublemakers." But a union can provide workers with a collective voice to communicate problems and grievances to management and to press for satisfactory resolutions.

A lower quit rate may give a firm a more experienced, more productive workforce. Over time, that might offset a part of the higher costs and reduced profitability associated with the union premium. Also, having fewer resignations might reduce the firm's recruitment, screening, and hiring costs. Additionally, reduced turnover may encourage employers to invest more in the training (and therefore the productivity) of their workers. If a worker quits or "exits" at the end of, say, a year's training, the employer will get no return from providing that training. Lower turnover increases the likelihood that the employer will receive a return on the training it provides, thereby increasing its willingness to upgrade the skills of its workforce. All these factors may increase the long-run productivity of the unionized labor force and therefore reduce the efficiency loss caused by the union wage premium.

APPENDIX SUMMARY

LO15.9 Relate who belongs to U.S. unions, the basics of collective bargaining, and the economic effects of unions.

About 8.4 million of the 14.8 million union members in 2011 belonged to unions affiliated with the AFL-CIO; another 5.5 million belonged to 7 unions loosely federated under the name Change to Win. The rest were members of other independent unions. About 11.8 percent of U.S. wage and salary workers in 2011 were union members, with government employees having the highest unionization rates. As an occupation, public school teachers have the highest rate of unionization—37 percent.

Union membership has declined as a percentage of the labor force and in absolute numbers in recent decades. Some of the key causes are structural changes such as the shift from manufacturing employment to service employment. Other causes include improved wages and working conditions in nonunion firms and increased managerial opposition to unions.

Collective bargaining determines the terms of union work agreements, which typically cover (*a*) union status and managerial prerogatives; (*b*) wages, hours, and working conditions; (*c*) control over job opportunities; and (*d*) grievance procedures. The bargaining process is governed by the National Labor Relations Act.

Union wages are on average about 15 percent higher than nonunion wages in comparable jobs. Restrictive union work rules, output losses from strikes, and labor misallocation from the union wage advantage are ways that unions may reduce efficiency, output, and productivity. The efficiency losses from unions may be partially offset in the long run by union productivity advances deriving from reduced labor turnover.

APPENDIX TERMS AND CONCEPTS

American Federation of Labor and the Congress of Industrial Organizations (AFL-CIO)
Change to Win
independent unions
unionization rate
collective bargaining
closed shop
union shop
agency shop
right-to-work laws
open shop
strike
lockout
National Labor Relations Act (NLRA)
National Labor Relations Board (NLRB)
voice mechanism
exit mechanism

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Which industries and occupations have the highest rates of unionization? Which the lowest? Speculate on the reasons for such large differences. **LO15.9**
2. What percentage of wage and salary workers are union members? Is this percentage higher, or is it lower, than in previous decades? Which of the factors explaining the trend do you think is most dominant? **LO15.9**
3. Explain how featherbedding and other restrictive work practices can reduce labor productivity. Why might strikes reduce the economy's output less than the loss of production by the struck firms? **LO15.9**
4. What is the estimated size of the union wage advantage? How might this advantage diminish the efficiency with which labor resources are allocated in the economy? Normally, labor resources of equal potential productivity flow from low-wage employment to high-wage employment. Why does that not happen to close the union wage advantage? **LO15.9**
5. Contrast the voice mechanism and the exit mechanism for communicating dissatisfaction. In what two ways do labor unions reduce labor turnover? How might such reductions increase productivity? **LO15.9**

APPENDIX REVIEW QUESTIONS

1. True or false. In the United States, unions have been gaining in membership and power for several decades. **LO15.9**
2. Suppose that you are president of a newly established local union about to bargain with an employer for the first time. List the basic areas you want covered in the work agreement. Why might you begin with a larger wage demand than you actually are willing to accept? What is the logic of a union threatening an employer with a strike during the collective bargaining process? Of an employer threatening the union with a lockout? What is the role of the deadline in encouraging agreement in collective bargaining? **LO15.9**
3. Look back at Figure 2. In the union sector, the union's ability to raise wages from W_n to W_u decreases total employment from N_1 to N_2. Thus $N_1 - N_2$ workers are displaced from the union sector and will seek employment in the nonunion sector. But suppose that wages in the nonunion sector cannot fall (perhaps because of a minimum wage law). Suppose, more specifically, that they are fixed at W_n in the nonunion sector. If the union and nonunion sectors are the only two sectors in the economy, how many workers will become unemployed because of the union's ability to raise wages in the union sector? (Hint: $N_1 - N_2$ in the union sector is the same number of workers as $N_3 - N_1$ in the nonunion sector.) **LO 15.9**
   a. N_1.
   b. N_2.
   c. N_3.
   d. $N_1 - N_2$.
   e. $N_1 + N_2$.
4. True or false. "To the extent that they succeed in their goals, unions only ever reduce productivity and efficiency." **LO15.9**

APPENDIX PROBLEMS

1. Suppose that a delivery company currently uses one employee per vehicle to deliver packages. Each driver delivers 50 packages per day, and the firm charges $20 per package for delivery. **LO15.9**

a. What is the MRP per driver per day?

b. Now suppose that a union forces the company to place a supervisor in each vehicle at a cost of $300 per supervisor per day. The presence of the supervisor causes the number of packages delivered per vehicle per day to rise to 60 packages per day. What is the MRP per supervisor per day? By how much per vehicle per day do firm profits fall after supervisors are introduced?

c. How many packages per day would each vehicle have to deliver in order to maintain the firm's profit per vehicle after supervisors are introduced?

d. Suppose that the number of packages delivered per day cannot be increased but that the price per delivery might potentially be raised. What price would the firm have to charge for each delivery in order to maintain the firm's profit per vehicle after supervisors are introduced?

2. Suppose that a car factory initially hires 1,500 workers at $30 per hour and that each worker works 40 hours per week. Then the factory unionizes, and the new union demands that wages be raised by 10 percent. The firm accedes to that request in collective bargaining negotiations but then decides to cut the factory's labor force by 20 percent due to the higher labor costs. **LO15.9**

a. What is the new union wage? How many workers does the factory employ after the agreement goes into effect?

b. How much in total did the factory's workers receive in wage payments each week before the agreement? How much do the factory's remaining workers receive in wage payments each week after the agreement?

c. Suppose that the workers who lose their jobs as a result of the agreement end up unemployed. By how much do the total wages received each week by the initial 1,500 workers (both those who continue to be employed at the factory and those who lose their jobs) change from before the agreement to after the agreement?

d. If the workers who lose their jobs as a result of the agreement end up making $15 per hour at jobs where they work 40 hours per week, by how much do the total wages received each week by the initial 1,500 workers change from before the agreement to after the agreement?

CHAPTER **16**

Rent, Interest, and Profit

Learning Objectives

LO16.1 Explain the nature of economic rent and how it is determined.

LO16.2 Define interest and explain how interest rates vary based on risk, maturity, loan size, and taxability.

LO16.3 Explain the loanable funds theory of interest rates.

LO16.4 Demonstrate how interest rates relate to the time-value of money.

LO16.5 Explain the role of interest rates in allocating capital, modulating R&D spending, and helping to determine the economy's total output of goods and services.

LO16.6 Relate why economic profits occur, and how profits, along with losses, allocate resources among alternative uses.

LO16.7 List the share of U.S. earnings received by each of the factors of production.

In Chapter 15, we focused on the wages and salaries paid by firms to obtain labor. Here our attention is focused on the rent, interest, and profit paid by firms to obtain, respectively, land, capital, and entrepreneurship. Our analysis will provide answers to numerous practical questions, including:

How do *land prices* and *land rents* get established, and why do they differ from property to property? For example, why do 20 acres of land in the middle of the Nevada desert sell for $5,000 while 20 acres along The Strip in Las Vegas command $500 million?

What determines *interest rates* and causes them to change? For instance, why were interest rates on 3-month bank certificates of deposit 1.3 percent in January 2003, 5.4 percent in June 2006, and 0.2 percent in November 2009? How does interest compound over time, and how does that compounding relate to the so-called present value and future value of a particular sum of money?

What are the sources of *profits* and *losses,* and why do they vary? For example, why did Walmart earn profits of nearly $16 billion in 2011 while financial firm MF Global went bankrupt after losing over $1.6 billion of its clients' money?

Economic Rent

LO16.1 Explain the nature of economic rent and how it is determined.

To most people, "rent" means the money paid for the use of an apartment or a room in a residence hall. To the business executive, "rent" is a payment made for the use of a factory building, machine, or warehouse facility owned by others. Such definitions of rent can be confusing and ambiguous, however. Residence hall room rent, for example, may include other payments as well: interest on money the university borrowed to finance the dormitory, wages for custodial services, utility payments, and so on.

Economists use "rent" in a much narrower sense. **Economic rent** is the price paid for the use of land and other natural resources that are completely fixed in total supply. As you will see, this fixed overall supply distinguishes rental payments from wage, interest, and profit payments.

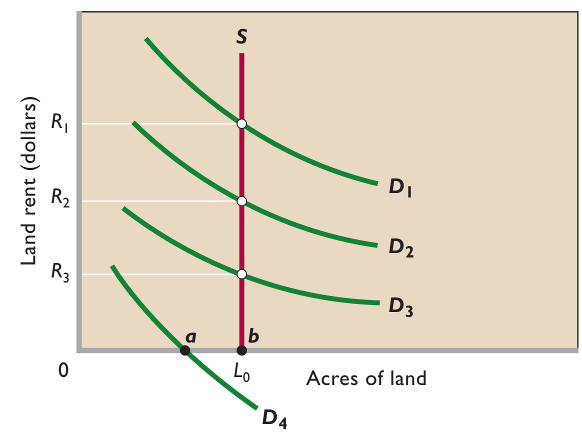
Let's examine this idea and some of its implications through supply and demand analysis. We first assume that all land has a single use, for example, growing wheat. We assume, too, that all land is of the same grade or quality, meaning that each arable (tillable) acre of land is as productive as every other acre. And we suppose that land is rented or leased in a competitive market in which many producers are demanding land and many landowners are offering land in the market.

In Figure 16.1, curve *S* represents the supply of arable land available in the economy as a whole, and curve D_2 represents the demand of producers for use of that land. As with all economic resources, the demand for land is a derived demand, meaning that the demand for land is derived from the demand for the products that land helps to produce. Demand curves such as D_2 reflect the marginal revenue product (MRP = MP × *P*) of land. The curve slopes downward because of diminishing returns (MP declines) and because, for producers as a group, additional units of land result in greater output and thus lower output prices (*P* is less).

Perfectly Inelastic Supply

The unique feature of our analysis is on the supply side. For all practical purposes the supply of land is perfectly inelastic (in both the short run and the long run), as reflected in supply curve *S*. Land has no production cost; it is a "free and nonreproducible gift of nature." The economy has only so much land, and that's that. Of course, within limits any parcel of land can be made more usable by clearing, drainage, and irrigation. But these are capital improvements and not changes in the amount of land itself. Moreover, increases in the usability of land affect only a small fraction of the total amount of land and do not change the basic fact that land and other nonrenewable natural resources are fixed in supply.

FIGURE 16.1 The determination of land rent. Because the supply *S* of land (and other natural resources) is perfectly inelastic, demand is the sole active determinant of land rent. An increase in demand from D_2 to D_1 or a decrease in demand from D_2 to D_3 will cause a considerable change in rent: from R_2 to R_1 in the first instance and from R_2 to R_3 in the second. But the amount of land supplied will remain at L_0. If demand is very weak (D_4) relative to supply, land will be a "free good," commanding no rent.



Equilibrium Rent and Changes in Demand

Because the supply of land is fixed, demand is the only active determinant of land rent. In this case, supply is passive. And what determines the demand for land? The factors we discussed in Chapter 14: the price of the products produced on the land, the productivity of land (which depends in part on the quantity and quality of the resources with which land is combined), and the prices of the other resources that are combined with land.

If demand is D_2, as we have suggested, the equilibrium rent will be R_2. The quantity of land L_0 that producers wish to rent will equal the quantity of land available (also L_0). But if the demand for land in Figure 16.1 increased from D_2 to D_1, land rent would rise from R_2 to R_1. On the other hand, if the demand for land declined from D_2 to D_3, land rent would fall from R_2 to R_3. Finally, if the demand for land were only D_4, land rent would be zero. In this situation, land would be a *free good*—a good for which demand is so weak relative to supply that an excess supply of it occurs even if the market price is zero. In Figure 16.1, we show this excess supply as distance $b - a$ at rent of zero. This essentially was the situation in the free-land era of U.S. history.

The ideas underlying Figure 16.1 help answer one of our chapter-opening questions. Land prices and rents are so high along the Las Vegas strip because the demand for that land is tremendous. It is capable of producing exceptionally high revenue from gambling, lodging, and entertainment. In contrast, the demand for isolated land in the middle of the desert is highly limited because very little revenue can be generated from its use. (It is an entirely different matter, of course, if gold can be mined from the land, as is true of some isolated parcels in Nevada!)

Productivity Differences and Rent Differences

So far we have assumed that all land is equally productive. That assumption is unrealistic because land varies widely in terms of productivity. As an example, differences in rainfall, soil quality, and other factors imply that while land in Kansas is excellently suited to wheat production, the sagebrush plains of Wyoming are much less well suited and the desert of Arizona is practically useless. Such productivity differences are reflected in resource demands and economic rents. Competitive bidding by producers will establish a high rent for highly productive Kansas land; less productive Wyoming land will command a much lower rent; and Arizona desert land may command no rent at all.

This process whereby differences in productivity lead to differences in rents can be understood graphically if we look at Figure 16.1 from a slightly different perspective. As before, assume that land can only be used for wheat production. But this time assume that there are four different plots of land. Each plot is of the same size L_0 but differs in productivity so that different marginal revenue products emerge when each plot of land is combined with identical amounts of labor, capital, and entrepreneurial talent.

These differences in marginal revenue products lead to four different demand curves: D_1, D_2, D_3, and D_4. D_1 is the highest demand curve because plot 1 has the highest productivity. D_4 is the lowest demand curve because plot 4 has the lowest productivity. When combined with supply curve S, the different demand curves yield different equilibrium rents: R_1, R_2, R_3, and R_4. The differences in rents mirror the differences in productivity so that plot 1 commands the highest rent, while plot 4 is so poor in quality that, given supply S, farmers won't pay anything to use it. It will be a free good because it is not sufficiently scarce in relation to its demand for it to command a price above zero.

As a final point, be aware that location itself can affect productivity and rent. Other things equal, renters will pay more for a unit of land that is strategically located with respect to materials, transportation, labor, and customers than they will for a unit of land whose location is remote from these things. Examples include the enormously high land prices near major ski resorts and the high price of land that contains oil beneath it.

Land Rent: A Surplus Payment

The supply of land is perfectly inelastic both in total and with respect to individual parcels of land. Whether land prices rise or fall, a nation will have the same total area to work with and individual plots of land will stay the same size.

The perfectly inelastic supply of land must be contrasted with the relatively elastic supply of nonland resources. Consider capital, which includes apartment buildings, fiber optic networks, and machinery. When the prices of these and other capital goods rise, entrepreneurs respond by increasing the production of capital goods. Conversely, a decline in capital goods prices results in reduced production of capital goods. As a result, the supply curves of nonland resources are normally upsloping, so that the prices paid to such resources provide an **incentive function.** A high price provides an incentive to offer more of the resource, whereas a low price prompts resource suppliers to offer less.

Not so with unimproved land. Rent serves no incentive function because both the total area of land in a nation as well as the sizes of individual plots of land will always stay exactly the same no matter what land prices are. As a result, economists consider land rents to be *surplus payments* that are not necessary to ensure that land is made available for economic use. From this perspective, the sum of all the land rents paid across a nation constitutes a giant surplus payment because it has no effect on the total supply of land in the nation. And in the same way, the individual land rents paid on particular plots of land are also surplus payments because they likewise have no effect on the sizes of those individual plots.

Land Ownership: Fairness versus Allocative Efficiency

If land is a gift of nature, costs nothing to produce, and would be available even without rental payments, why should rent be paid to those who just happen to be landowners? Socialists have long argued that all land rents are unearned incomes because the act of owning land and renting it out to others produces nothing of value in and of itself. They urge that land should be nationalized (owned by the state) so that any payments for its use can be put to work by the government to further the well-being of the entire population.

Opponents of land nationalization argue that private land ownership allows Adam Smith's "invisible hand" to work its magic in terms of allocating scarce land resources to their best possible uses. In a nation where land is privately owned and rents are charged for the use of land, individuals and firms are forced to consider opportunity costs when deciding whether to secure the use of a particular piece of land. This gives them an incentive to allocate each piece of land to its highest-value use. In particular, renters will only allocate land to uses that generate enough revenue to both pay the rent and cover all other costs, including a normal profit.

Private land ownership and having to pay market-determined land rents also aid economic growth and development because as consumer tastes change and as new technologies are developed, the best uses to which particular pieces of land can be put also change. These changing opportunity costs are reflected in land rents, whose changing values thereby help to reallocate land from lower-value uses to higher-value uses as the economy evolves—thus, the often-heard remark, "The land was just too valuable for its previous use."

By contrast, if land were nationalized, government planners would have a difficult time assigning each piece of land to its best possible use and adjusting its use with changing circumstances without the guidance about opportunity costs provided by market-determined rents.

Along those lines, it is important to be clear that while economic rents are surplus payments when viewed from the perspective of society as a whole, they are most definitely costs to individual people and individual firms. To see why this is true, recall that because land is a free gift of nature, there is no cost to society as a whole for obtaining the current supply of land. Thus, economic rents are, from the perspective of society, surplus payments because they have no effect on land supply. But individuals must pay economic rents because such rents determine how society's fixed supply of land is allocated among competing potential uses. Those who are willing and able to pay the market rent get to use the land, while those who are unwilling or unable to pay the market rent do not. Put slightly differently: Economic rents do not cause land to be supplied—they cause land to be directed.

Application: A Single Tax on Land

In the United States, criticism of rental payments produced the **single-tax movement,** which gained significant support in the late nineteenth century. Spearheaded by Henry George's provocative book *Progress and Poverty* (1879), supporters of this reform movement held that economic rent could be heavily taxed without diminishing the available supply of land or reducing the efficiency with which it is allocated.

Henry George's Proposal George observed that, as population grew and the Western frontier closed, landowners enjoyed larger and larger rents (or lease incomes) from their landholdings. That increase in rents was the result of a growing demand for a resource whose supply was perfectly inelastic. Some landlords were receiving fabulously high incomes, not through any productive effort but solely through their ownership of highly prized land. George insisted that these increases in land rent belonged to society at large. Consequently, he argued that land rents should be heavily taxed and that the revenue generated by land taxes be spent for public uses. In seeking popular support for his ideas on land taxation, George proposed that a tax on rental income be the *only* tax levied by government.

George's case for taxing rental income was based not only on equity or fairness but also on efficiency. First, he wished for land to remain in private hands so that the "invisible hand" would guide private landowners to allocate their land to its best possible use. Allocative efficiency would still be achieved because the most profitable use for a particular piece of land before it is taxed remains the most profitable use for that land after it is taxed. Thus, landowners would not have any incentive to shift the use of their land from one activity to another just because it was taxed. In addition, landlords would not withdraw land from production when the tax was imposed because doing so would mean no rental income at all. Because some rental income, no matter how small, is better than no rental income, landlords would continue to supply land.

George further argued that, in addition to maintaining allocative efficiency for land, his single tax would also be much better for society in terms of productive efficiency. His argument was based on the fact that because land is in fixed supply, it does not have an incentive function. Thus, when taxed, the quantity of land supplied does not change. By contrast, because other resources have incentive functions, taxing them would likely reduce their supply—something that would cause an underproduction of output and, consequently, productive inefficiency. For instance, if workers responded to a tax on labor by supplying less labor, output would decline. Similarly, if a property tax on buildings caused builders to supply fewer buildings, there would be less capital available to produce output and, hence, less output. But because land is in fixed supply, no such reductions in output need be feared when taxing land. Thus, if only land were taxed, society would never have to worry about taxes causing productive inefficiency by reducing output.

Criticisms The single tax on land has few remaining advocates. Critics of the idea have pointed out that:

- Current levels of government spending are such that a land tax alone would not bring in enough revenue.
- Most income payments consist of a mixture of interest, rent, wages, and profits. So in practice it would be difficult to isolate how much of any specific income payment is actually derived from rent.
- So-called unearned income accrues to many people other than landowners. For example, consider the capital-gains income received by someone who many years ago purchased or inherited stock that now delivers hefty dividend payments. Is such income more "earned" than the rental income of the landowner?
- Historically, a piece of land is likely to have changed ownership many times. It would therefore be highly unfair to impose a heavy tax on recent buyers who paid a high price for their land and thus did not gain in any way from previous price increases.

QUICK REVIEW 16.1

- Rent is the price paid for unimproved land whose supply is perfectly inelastic.
- Differences in rent arise from differences in land productivity and location.
- Rent is socially useful because it puts an opportunity cost on land parcels, thereby helping to allocate each parcel of land to its best possible use.
- Henry George proposed a single tax on land rent, noting that rent was a surplus payment and that taxing it would not distort land allocation.
- Contemporary economists dismiss George's single-tax proposal as inadequate, impractical, and unfair.

Interest

LO16.2 Define interest and explain how interest rates vary based on risk, maturity, loan size, and taxability.

Interest is the price paid for the use of money. It can be thought of as the amount of money that a borrower must pay a lender for the use of the lender's money over some period of time. As an example, a borrower might be required to pay $100 of interest for the use of $1,000 for one year.

Because borrowers pay for loans of money with money, interest can be stated as a percentage of the amount of money borrowed rather than as a dollar amount. This is useful because it is far less clumsy to say that interest is "12 percent annually" than to say that interest is "$120 per year per $1,000."

Stating interest as a percentage also makes it much easier to compare the interest paid on loans involving different amounts of money. By expressing interest as a percentage, we can immediately compare an interest payment of, say, $432 per year per $2,880 with one of $1,800 per year per $12,000. Both interest payments are 15 percent per year, which is not obvious from the actual dollar figures.

And to make things even simpler, an interest payment of 15 percent per year can also be referred to as a 15 percent interest rate.

Money Is Not a Resource

When considering why borrowers are willing to pay interest for the right to use borrowed money, it is important to remember that money is not itself an economic resource. Whether money comes in the form of coins, paper currency, or checking accounts, you cannot directly produce any goods and services with it.

Thus, borrowers do not value money for its own sake. Rather, they value money because of what it can purchase. Individuals and households are willing to pay interest to borrow for consumption spending because they would rather consume certain goods and services sooner rather than later. And businesses are willing to pay interest because the money that they borrow can be used to expand their businesses and increase their profits. In particular, borrowed money can be used to fund the acquisition of capital goods such as computers, machinery, and warehouses.

Interest Rates and Interest Income

The interest rate on money loans determines the *interest income* earned by households for providing capital to firms. This is true because firms have the choice of either leasing capital from households or purchasing their own capital. Because businesses have this option, households wishing to lease their capital to businesses cannot charge more for the use of their capital than what businesses would have to pay in terms of interest payments to borrow the money needed to purchase their own capital.

As an example, consider a custom T-shirt shop that needs a $10,000 embroidering machine to expand production. If the owners of the shop can borrow the money to buy such a machine at an interest rate of 8 percent per year, then anyone wishing to lease them an identical machine could charge them no more than $800 per year for it (since $800 is how much per year the shop would have to pay in interest to borrow $10,000 at an 8 percent interest rate.)

Range of Interest Rates

For convenience, economists often speak in terms of a single interest rate. However, there are actually a number of interest rates in the economy. Table 16.1 lists several interest rates often referred to in the media. On November 19, 2012, these rates ranged from 0.09 to 14.58 percent. Why the differences?

- ***Risk*** Loans to different borrowers for different purposes carry varying degrees of risk. The greater the chance that a borrower will not repay his loan, the higher the interest rate the lender will charge to compensate for that risk.
- ***Maturity*** The time length of a loan, or its *maturity* (when it needs to be paid back), also affects the interest rate. Other things equal, longer-term loans usually command higher interest rates than shorter-term loans. This is true because one function of interest rates is to compensate lenders for the inconvenience and potential financial sacrifices involved with forgoing alternative uses of their money until their loans are repaid. Longer-term loans must offer higher interest rates to compensate lenders for having to forgo alternative opportunities for longer periods of time.
- ***Loan size*** If there are two loans of equal maturity and risk, the interest rate on the smaller of the two loans usually will be higher. The administrative costs of issuing a large loan and a small loan are about the same in dollars, but the cost is greater *as a percentage* of the smaller loan.
- ***Taxability*** Interest on certain state and municipal bonds is exempt from the federal income tax. Because lenders are interested in their after-tax rate of interest, the bonds issued by state and local governments can attract lenders even though they pay lower before-tax interest rates than other bonds of similar maturity and risk. Consider a lender who expects to pay a 35 percent federal income tax on any taxable interest payments. She may prefer a 5 percent interest rate on a tax-exempt municipal bond to a 6 percent interest rate on a taxable corporate bond because, after paying taxes on that taxable corporate bond, she will be left with an after-tax return of only 3.9 percent.

TABLE 16.1 **Selected Interest Rates, November 19, 2012**

Type of Interest Rate	Annual Percentage
20-year Treasury bond rate (interest rate on federal government security used to finance the public debt)	2.34%
90-day Treasury Bill rate (interest rate on federal government security used to finance the public debt)	0.09
Prime interest rate (interest rate used as a reference point for a wide range of bank loans)	3.25
30-year mortgage rate (fixed-interest rate on loans for houses)	3.34
4-year automobile loan rate (interest rate for new autos by automobile finance companies)	4.17
Tax-exempt state and municipal bond rate (interest rate paid on a low-risk bond issued by a state or local government)	3.41
Federal funds rate (interest rate on overnight loans between banks)	0.16
Consumer credit card rate (interest rate charged for credit card purchases)	14.58

Sources: Federal Reserve, **www.federalreserve.gov**, and Bankrate.com, **www.bankrate.com**.

Pure Rate of Interest

When economists and financial specialists talk of "the" interest rate, they typically have in mind the **pure rate of interest.** This is the hypothetical interest rate that would serve purely and solely to compensate lenders for their willingness to patiently forgo alternative consumption and investment opportunities until their money is repaid.

The pure rate of interest is best approximated by the interest rates of long-term, virtually riskless securities, such as the 20-year Treasury bonds issued by the U.S. federal government. Because such bonds involve minimal risk and negligible administrative costs, the interest that they pay can be thought of as compensating purely and solely for the use of money over an extended period of time. In November 2012, the pure rate of interest in the United States was 2.34 percent.

Loanable Funds Theory of Interest Rates

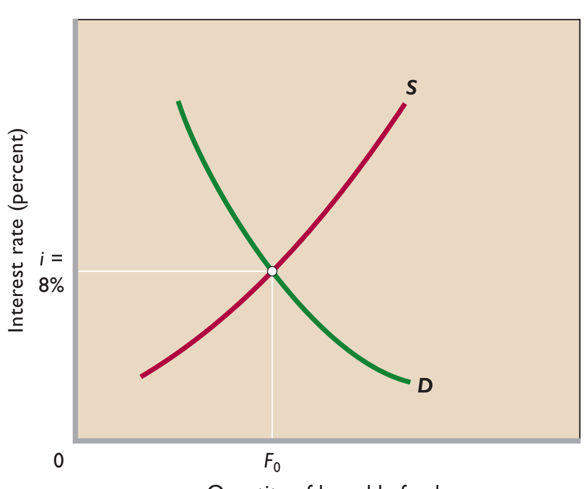
LO16.3 Explain the loanable funds theory of interest rates.

Because macroeconomics deals with the entire economy, it typically focuses on the pure rate of interest and assumes that it is determined by the total supply and demand for money in the economy. But because our present focus is on microeconomics, we will focus on a more micro-based theory of interest rates. By doing so, we will be able to explain why the interest rates on different types of loans vary so greatly, as in Table 16.1.

The **loanable funds theory of interest** explains the interest rate on any particular type of loan in terms of the supply of and demand for *funds available for lending* in the loanable funds market that exists for that particular type of loan. As Figure 16.2 shows, the equilibrium interest rate (here, 8 percent) on a particular type of loan is the rate at which the quantities of loanable funds supplied and demanded are equal for that type of loan.

To gain a deeper understanding of the loanable funds theory of interest, let's focus on a simplified lending market. First, assume that, for a particular type of loan, households or consumers are the sole suppliers of loanable funds, while businesses are the sole demanders of loanable funds. Also assume that lending occurs directly between households and businesses so that there are no financial institutions acting as intermediaries.

FIGURE 16.2 A loanable funds market. The upsloping supply curve S for loanable funds in a specific lending market reflects the idea that at higher interest rates, households will defer more of their present consumption (save more), making more funds available for lending. The downsloping demand curve D for loanable funds in such a market indicates that businesses will borrow more at lower interest rates than at higher interest rates. At the equilibrium interest rate (here, 8 percent), the quantities of loanable funds lent and borrowed are equal (here, F_0 each).



Supply of Loanable Funds

The supply of loanable funds in our simplified lending market is represented by curve S in Figure 16.2. Its upward slope indicates that households will make available a larger quantity of funds at high interest rates than at low interest rates. Most people prefer to use their incomes to purchase goods and services *today*, rather than delay purchases to sometime in the *future*. For people to delay consumption and increase their saving, they must be "bribed" or compensated by an interest payment. The larger the amount of that payment, the greater the deferral of household consumption and thus the greater the amount of money made available for loans.

Demand for Loanable Funds

Businesses borrow loanable funds primarily to add to their stocks of capital goods, such as new plants or warehouses, machinery, and equipment. Assume that a firm wants to buy a machine that will increase output and sales such that the firm's total revenue will rise by $110 for the year. Also assume that the machine costs $100 and has a useful life of just 1 year. Comparing the $10 earned with the $100 cost of the machine, we find that the expected rate of return on this investment is 10 percent (= $10/$100) for the 1 year.

To determine whether the investment would be profitable and whether it should be made, the firm must compare the interest rate—the price of loanable funds—with the 10 percent expected rate of return. If funds can be borrowed at some rate less than the rate of return, say, at 8 percent, as in Figure 16.2, then the investment is profitable and should be made. But if funds can be borrowed only at an interest rate above the 10 percent rate of return, say, at 14 percent, the investment is unprofitable and should not be made.

Why is the demand for loanable funds downsloping, as in Figure 16.2? At higher interest rates fewer investment projects will be profitable and therefore a smaller quantity of loanable funds will be demanded. At lower interest rates, more investment projects will be profitable and therefore more loanable funds will be demanded. Indeed, as we have just seen, purchasing the $100 machine is profitable if funds can be borrowed at 8 percent but not if the firm must borrow at 14 percent.

ORIGIN OF THE IDEA

O16.1

Interest rates

Extending the Model

There is a loanable funds market for nearly every type of loan in the economy. The best known are the markets for corporate and government bonds. But there are also loanable funds markets for student loans, home mortgages, and car loans. Each type of loan ends up with its own equilibrium interest rate determined by the demand and supply for loanable funds in its particular market.

We now extend the simple loanable funds model to make it more realistic and better able to capture the diversity of these many lending markets.

Financial Institutions Households rarely directly lend their savings to the businesses that are borrowing funds for investment. Instead, they place their savings in banks (and other financial institutions). The banks pay interest to savers in order to attract loanable funds and in turn lend those funds to businesses. Businesses borrow the funds from the banks, paying them interest for the use of the money. Financial institutions profit by charging borrowers higher interest rates than the interest rates they pay savers. Both interest rates, however, are based on the supply of and demand for loanable funds in their respective markets.

Changes in Supply Anything that causes households to be thriftier will prompt them to save more at each interest rate, shifting the supply curve rightward. For example, if interest earned on savings were to be suddenly exempted from taxation, we would expect the supply of loanable funds to increase and the equilibrium interest rate to decrease.

Conversely, a decline in thriftiness would shift the supply-of-loanable-funds curve leftward and increase the equilibrium interest rate. Illustration: If the government expanded social insurance to cover the costs of hospitalization, prescription drugs, and retirement living more fully, the incentive of households to save might diminish.

Changes in Demand On the demand side, anything that increases the rate of return on potential investments will increase the demand for loanable funds. Let's return to our earlier example, where a firm would receive additional revenue of $110 by purchasing a $100 machine and, therefore, would realize a 10 percent return on investment. What factors might increase or decrease the rate of return? Suppose a technological advance raised the productivity of the machine such that the firm's total revenue increased by $120 rather than $110. The rate of return would then be 20 percent, not 10 percent. Before the technological advance, the firm would have demanded zero loanable funds at, say, an interest rate of 14 percent. But now it will demand $100 of loanable funds at that interest rate, meaning that the demand curve for loanable funds has been shifted to the right.

Similarly, an increase in consumer demand for the firm's product will increase the price of its product. So even though the productivity of the machine is unchanged, its potential revenue will rise from $110 to perhaps $120, increasing the firm's rate of return from 10 to 20 percent. Again the firm will be willing to borrow more than previously at our presumed 8 or 14 percent interest rate, implying that the demand curve for loanable funds has shifted rightward. This shift in demand increases the equilibrium interest rate.

Conversely, a decline in productivity or in the price of the firm's product would shift the demand curve for loanable funds leftward, reducing the equilibrium interest rate.

Other Participants We must recognize that participation in many loanable funds markets may go well beyond our simplification of households as suppliers of funds and businesses as demanders of funds. For example, while households are suppliers of loanable funds, many are also demanders of such funds. Households borrow to finance expensive purchases such as housing, automobiles, furniture, and household appliances. Governments also are on the demand side of a loanable funds market when they borrow to finance budgetary deficits. And businesses that have revenues in excess of their current expenditures may offer some of those revenues in various loanable funds markets. Thus, like households, businesses operate on both the supply and the demand sides of various loanable funds markets.

Finally, in addition to gathering and making available the savings of households, banks and other financial institutions also increase funds through the lending process and decrease funds when the money that is used to pay off loans is retained by the banks rather than being lent out again to other borrowers. The Federal Reserve (the nation's central bank) controls the amount of this bank activity and thus influences a wide variety of interest rates.

This fact helps answer one of our chapter-opening questions: Why did the interest rate on 3-month certificates of deposit in the United States fall from 5.4 percent in June 2006 to only 0.2 percent in November 2009? There are two reasons: (1) The demand for loanable funds sharply declined because businesses reduced their desire to purchase more capital goods and (2) the Federal Reserve, fighting recession and sluggish recovery, took monetary actions that greatly increased the supply of loanable funds. In contrast, between 2003 and 2006 the Federal Reserve restricted the growth of loanable funds. Because the demand for loanable funds increased more rapidly than the supply of loanable funds, interest rates such as those on 3-month certificates of deposit rose. As indicated in the chapter opening, that rate increased from 1.3 percent in January 2003 to 5.4 percent in June 2006.

Time-Value of Money

LO16.4 Demonstrate how interest rates relate to the time-value of money.

Interest is central to understanding the **time-value of money**—the idea that a specific amount of money is more valuable to a person the sooner it is obtained. To see where money's time value comes from, suppose that you could choose between being paid $1,000 today or $1,000 in a year. The fact that $1,000 received today can be invested at interest and grow into more than $1,000 in a year implies that it is better to receive $1,000 today than $1,000 in a year. By how much is it better? By the amount of interest that can be gained over the course of the year. The higher the interest rate, the greater the time-value of money.

In addition to giving money its time value, the fact that money can be invested to earn interest also implies a way in which a given amount of money today can be thought of as being equivalent to a larger amount of money in the future and how a future amount of money can be thought of as being equivalent to a smaller amount of money today. We explore this idea next.

Compound Interest

Compound interest is the total interest that cumulates over time on money that is placed into an interest-bearing account. Table 16.2 helps us explain compound interest, as well as the related ideas of future value and present value. Suppose that Max places $1,000 into an interest-bearing account at 10 percent interest with the intent to let the *principal* (the initial deposit) and interest compound for 3 years. The first row of each column shows the beginning period sum; the second column shows the yearly computation as to how that sum grows, given a particular interest rate. That growth is found by multiplying the dollar amount at the beginning of each year by $1 + i$, where i is the interest rate expressed as a decimal.

In year 1 the 10 percent interest rate increases the money in the account from $1,000 to $1,100 (= $1,000 × 1.10). So, as shown in Column 3, total interest is $100. Column 4 simply lists the $1,100 again but reinforces that this amount consists of the original principal plus the total interest. Similarly, in year 2, the $1,100 now in the account grows to $1,210 (= $1,100 × 1.10) because $110 of new interest accrues on the $1,100. At the end of year 2, the principal remains $1,000, but the total interest is $210 and the total amount in the account is $1,210. Interest in year 3 is $121 and total interest rises to $331. After this $331 of total interest is added to the $1,000 principle, the accumulation is $1,331. As shown in Column 3, compound interest builds and builds over time.

Future Value and Present Value

Now note from Table 16.2 that we can look at the time-value of money in two distinct ways. **Future value** is the amount to which some current amount of money will grow as interest compounds over time. In our table, the future value (FV) of $1,000 today at 10 percent interest is $1,331 three years from now. Future value is always forward-looking.

But we can just as easily look backward from the end value of $1,331 and ask how much that amount is worth today, given the 10 percent interest rate. **Present value** is today's value of some amount of money to be received in the future. In terms of the table, the present value (PV) of $1,331 is $1,000. Here, FV is "discounted" by three years at 10 percent to remove the $331 of compounded interest and therefore to obtain PV. (We will defer explaining the discounting procedure to our chapter on financial economics in the macro portion of *Economics*. But if you are interested in the mathematics, see the footnote below.)[1]

With any positive interest rate (and assuming no inflation), a person would prefer to receive $1,000 today rather than $1,000 at some time in the future. The higher the interest rate, the greater is the *future value* of a specific amount of money today. To confirm, substitute a 20 percent interest rate for the 10 percent rate in Table 16.2 and rework the analysis. Finally, you should know that the analysis presented in the table is extendable to any number of years.

[1]The mathematics is as follows:

$$FV = PV(1 + i)^t \quad \text{and} \quad PV = \frac{FV}{(1 + i)^t}$$

where i is the interest rate and t is time, here the number of years of compounding.

TABLE 16.2 Compound Interest, Future Value, and Present Value, 10 Percent Interest Rate

(1) Beginning Period Value	(2) Computation	(3) Total Interest	(4) End Period Value
$1,000 (Year 1)	$1,000 × 1.10 = $1,100	$100	$1,100 (= $1,000 + $100)
$1,100 (Year 2)	$1,100 × 1.10 = $1,210	$210 (= $100 + $110)	$1,210 (= $1,000 + $210)
$1,210 (Year 3)	$1,210 × 1.10 = $1,331	$331 (= $100 + $110 + $121)	$1,331 (= $1,000 + $331)

CONSIDER THIS . . .

That Is Interest

The following story told by economist Irving Fisher (1867–1947) helps illustrate the time-value of money.

In the process of a massage, a masseur informed Fisher that he was a socialist who believed that "interest is the basis of capitalism and is robbery." Following the massage, Fisher asked, "How much do I owe you?"

The masseur replied, "Thirty dollars."

"Very well," said Fisher, "I will give you a note payable a hundred years hence. I suppose you have no objections to taking this note without any interest. At the end of that time, you, or perhaps your grandchildren, can redeem it."

"But I cannot afford to wait that long," said the masseur.

"I thought you said that interest was robbery. If interest is robbery, you ought to be willing to wait indefinitely for the money. If you are willing to wait ten years, how much would you require?"

"Well, I would have to get more than thirty dollars."

His point now made, Fisher replied, "That is interest."*

*Irving Fisher, as quoted in Irving Norton Fisher, *My Father Irving Fisher* (New York: Comet Press Books, 1956), p. 77.

The time-value of money is an important concept. For example, it helps explain the optimal timing of natural resource extraction (Chapter 17). It also is critical to the entire field of financial economics (Chapter 35 of *Economics* or Chapter 17 of the macro-split version of *Economics*). In our present chapter, our goal is simply to stress that *money has time value because of the potential for compound interest*.

Role of Interest Rates

LO16.5 Explain the role of interest rates in allocating capital, modulating R&D spending, and helping to determine the economy's total output of goods and services.

We have already explained that interest rates on money loans determine the interest incomes earned by the owners of capital goods. This fact implies that interest rates are the critical prices determining both the *level* and *composition* of new investments in capital goods as well as the amount of research and development (R&D) spending in the economy.

Interest and Total Output

Lower equilibrium interest rates encourage businesses to borrow more for investment, other things equal. As a result, total spending in the economy rises, and if the economy has unused resources, so does total output. Conversely, higher equilibrium interest rates discourage businesses from borrowing for investment, thereby reducing investment and total spending. Such a decrease in spending may be desirable if an economy is experiencing inflation.

The Federal Reserve often manages interest rates to try to expand investment and output, on the one hand, or to reduce investment and inflation, on the other. It affects interest rates by changing the supply of money. Increases in the money supply increase the supply of loanable funds, causing equilibrium interest rates to fall. This boosts investment spending and expands the economy. In contrast, decreases in the money supply decrease the supply of loanable funds, boosting equilibrium interest rates. As a result, investment is constrained and so is the economy.

Interest and the Allocation of Capital

Prices are rationing devices. And interest rates are prices. Thus, when it comes to allocating capital in the economy, the interest rates charged on investment loans ration the available supply of loanable investment funds to investment projects that have expected rates of return at or above the interest rate cost of the borrowed funds.

If, say, the computer industry expects to earn a return of 12 percent on the money it invests in physical capital and it can secure the required funds at an interest rate of 8 percent, it can borrow and expand its physical capital. If the expected rate of return on additional capital in the steel industry is only 6 percent, that industry will find it unprofitable to expand its capital at 8 percent interest. The interest rate allocates money, and ultimately physical capital, to the industries in which it will be most productive and therefore most profitable. Such an allocation of capital goods benefits society.

But the interest rate on investment loans does not perfectly ration capital to its most productive uses. Large oligopolistic borrowers may be better able than competitive borrowers to pass interest costs on to consumers because they can change prices by controlling output. Also, the size, prestige, and monopsony power of large corporations may help them obtain funds on more favorable terms than can smaller firms, even when the smaller firms have similar rates of profitability.

Interest and R&D Spending

In Web Chapter 13, we pointed out that, similar to an investment decision, a decision on how much to spend on R&D depends on the cost of borrowing funds in relationship to the expected rate of return. Other things equal, the lower the interest rate and thus the lower the cost of borrowing funds for R&D, the greater is the amount of R&D spending that is potentially profitable. Low interest rates encourage R&D spending; high interest rates discourage it.

Also, the interest rate allocates R&D funds to firms and industries for which the expected rate of return on R&D is the greatest. Ace Microcircuits may have an expected rate of return of 16 percent on an R&D project, while Glow Paints has only a 2 percent expected rate of return on an R&D project. With the interest rate at 8 percent, loanable funds will flow to Ace, not to Glow. Society will benefit by having R&D spending allocated to projects that have high enough expected rates of return to justify using scarce resources for R&D rather than for other purposes.

Nominal and Real Interest Rates

This discussion of the role of interest in investment decisions and in R&D decisions assumes that there is no inflation. If inflation exists, we must distinguish between nominal and real interest rates, just like we needed to distinguish between nominal and real wages in Chapter 15. The **nominal interest rate** is the rate of interest expressed in dollars of current value. The **real interest rate** is the rate of interest expressed in purchasing power—dollars of inflation-adjusted value. (For a comparison of nominal interest rates on bank loans in selected countries, see Global Perspective 16.1.)

Example: Suppose the nominal interest rate and the rate of inflation are both 10 percent. If you borrow $100, you must pay back $110 a year from now. However, because of 10 percent inflation, each of these 110 dollars will be worth 10 percent less. Thus, the real value or purchasing power of your $110 at the end of the year is only $100. In inflation-adjusted dollars you are borrowing $100 and at year's end you are paying back $100. While the nominal interest rate is 10 percent, the real interest rate is zero. We determine the real interest rate by subtracting the 10 percent inflation rate from the 10 percent nominal interest rate.

It is the real interest rate, not the nominal rate, that affects investment and R&D decisions.

Application: Usury Laws

A number of states have passed **usury laws,** which specify a maximum interest rate at which loans can be made. Such rates are a special case of *price ceilings*, discussed in Chapter 3. The purpose of usury laws is to hold down the interest cost of borrowing, particularly for low-income borrowers. ("Usury" simply means exorbitant interest.)

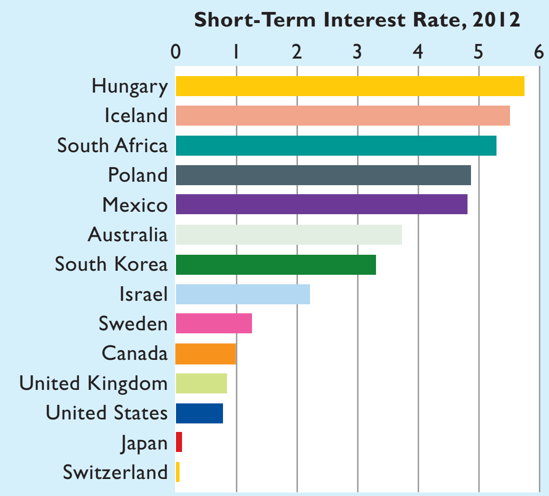
Figure 16.2 helps us assess the impact of such legislation. The equilibrium interest rate there is 8 percent, but suppose a usury law specifies that lenders cannot charge more than 6 percent. The effects are as follows:

- ***Nonmarket rationing*** At 6 percent, the quantity of loanable funds demanded exceeds the quantity supplied: There is a shortage of loanable funds. Because the market interest rate no longer can ration the available loanable funds to borrowers, lenders (banks) have to do the rationing. We can expect them to make loans only to the most creditworthy borrowers (mainly wealthy, high-income people), thus defeating the goal of the usury law. Low-income, riskier borrowers are

GLOBAL PERSPECTIVE 16.1

Short-Term Nominal Interest Rates, Selected Nations

These data show the short-term nominal interest rates (percentage rates on 3-month loans) in various countries in 2012. Because these are nominal rates, much of the variation reflects differences in rates of inflation. But differences in central bank monetary policies and default risk also influence the variation.



Source: OECD Economic Outlook, Organization for Economic Cooperation and Development, **www.oecd.org**.

ORIGIN OF THE IDEA

O16.2

Usury

excluded from the market and may be forced to turn to loan sharks who charge illegally high interest rates.

- ***Gainers and losers*** Creditworthy borrowers gain from usury laws because they pay below-market interest rates. Lenders (ultimately bank shareholders) are losers because they receive 6 percent rather than 8 percent on each dollar lent.
- ***Inefficiency*** We have just seen how the equilibrium interest rate allocates money to the investments and the R&D projects whose expected rates of return are greatest. Under usury laws, funds are much less likely to be allocated by banks to the most productive projects. Suppose Mendez has a project so promising she would pay 10 percent for funds to finance it. Chen has a less-promising investment, and he would be willing to pay only 7 percent for financing. If the market were rationing funds, Mendez's highly productive project would be funded and Chen's would not. That allocation of funds would be in the interest of both Mendez and society. But with a 6 percent usury rate, Chen may get to the bank before Mendez and receive the loanable funds at 6 percent. So Mendez may not get funded. Legally controlled interest rates may thus inefficiently ration funds to less-productive investments or R&D projects.

QUICK REVIEW 16.2

- Interest is the price paid for the use of money and determines the interest income earned by households for providing capital to firms.
- The range of interest rates is influenced by risk, maturity, loan size, and taxability.
- In the loanable funds model, the equilibrium interest rate is determined by the demand for and supply of loanable funds.
- The time-value of money is the idea that \$1 can be converted into more than \$1 of future value through compound interest and therefore that \$1 to be received sometime in the future has less than \$1 of present value.
- Interest rates on investment loans affect the total level of investment and therefore the levels of total spending and total output; they also allocate money and real capital to specific industries and firms. Similarly, interest rates also affect the level and composition of R&D spending.
- Usury laws that establish an interest-rate ceiling below the market interest rate may (a) deny credit to low-income people, (b) subsidize high-income borrowers and penalize lenders, and (c) diminish the efficiency with which loanable funds are allocated to investment and R&D projects.

Economic Profit

LO16.6 Relate why economic profits occur, and how profits, along with losses, allocate resources among alternative uses.

Recall from previous chapters that economists define profit narrowly. To accountants, "profit" is what remains of a firm's total revenue after it has paid individuals and other firms for the materials, capital, and labor they have supplied to the firm. To the economist, these "accounting profits" overstate profit. The reason is that the accountant's view of profit considers only **explicit costs:** payments made by the firm to outsiders. It ignores **implicit costs:** the monetary income the firm sacrifices when it uses resources that it owns, rather than supplying those resources to the market. The economist considers implicit costs to be opportunity costs that must be accounted for in determining profit. **Economic,** or **pure, profit** is what remains after all costs—both explicit and implicit costs, the latter including a normal profit—have been subtracted from a firm's total revenue. Economic profit may be either positive or negative (a loss).

Entrepreneurship and Profit

Economic profit flows to individuals to motivate them to provide the economic resource known as entrepreneurship. For illustration, let's suppose that two entrepreneurs establish a start-up company called Upside and that they are the only owners. They do not incorporate or borrow and instead use their own funds to finance the firm.

As discussed in previous chapters, the resource that these entrepreneurs provide is clearly not labor. Typical workers simply complete assigned tasks and engage in routine activities. For their labor inputs, they are compensated with wages and salaries.

Entrepreneurs, by contrast, make nonroutine decisions that involve substantial financial risk. Among other things, they (1) decide their firm's strategy for combining land, labor, and capital to produce a good or service; (2) decide whether and how to develop new products and new production processes; and (3) personally bear the financial risks associated with the success or failure of their respective firms.

With regard to bearing those financial risks, it is crucial to understand that a firm's entrepreneurs are its *residual claimants*, meaning that they only receive whatever residual revenue—if any—remains after all the other factors of production have been paid. As residual claimants, the entrepreneurs at Upside receive whatever accounting profit or accounting loss the firm generates. Thus, the financial risks of running the firm are borne by its two entrepreneurs. If Upside loses money, it loses *their* money.

Insurable and Uninsurable Risks

As residual claimants, entrepreneurs face financial risks. These fall into two categories. **Insurable risks** are those risks for which it is possible to buy insurance from an insurance company, while **uninsurable risks** are those risks for which it is not possible to buy insurance from an insurance company.

Individuals and firms who purchase insurance policies (contracts) from an insurance company are referred to as policyholders. In exchange for an annual premium (fee), the policyholders obtain the insurance company's guarantee to reimburse them for any financial losses caused by any of the risks covered under the insurance contract. In order to be able to keep that promise, the insurance company must be able to collect enough in premiums from its policyholders in the present to be able to fully reimburse the ones who eventually suffer losses in the future. This is only possible with risks like fire, flood, theft, accident, and death whose frequencies of occurrence can be predicted with relative accuracy. Risks whose frequencies of occurrence cannot be predicted with accuracy are uninsurable.

Sources of Uninsurable Risks

Because insurance is available for insurable risks, entrepreneurs only have to deal with uninsurable risks whose frequency of occurrence cannot be predicted with any accuracy. In practice, these are the result of uncontrollable and unpredictable changes in demand and supply that either reduce revenues or increase costs. These uninsurable risks fall into four main categories:

- ***Changes in the general economic environment*** An economy-wide downturn in business (a recession) can lead to greatly reduced demand, sales, and revenues, and thus to business losses. An otherwise prosperous firm may experience substantial losses through no fault of its own.
- ***Changes in the structure of the economy*** Consumer tastes, technology, resource availability, and prices change unpredictably in the real world, bringing changes in production costs and revenues. For example, an airline earning an economic profit one year may sustain substantial losses the next year as the result of a significant increase in the price of jet fuel.
- ***Changes in government policy*** A newly instituted regulation, the removal of a tariff, or a change in national defense policy may significantly alter the costs and revenues of the affected industry and firms.
- ***New products or production methods pioneered by rivals*** Any firm can suddenly find itself losing sales and revenue to popular new products brought out by rivals. Similarly, a firm may suddenly find itself having to sell its product at a loss if rival firms cut their prices after figuring out a lower-cost way to make the same product.

Profit as Compensation for Bearing Uninsurable Risks

Economists list entrepreneurship as its own economic resource—separate from land, labor, and capital—because it is not possible to run a business without somebody being willing to undertake and live with uninsurable risks. Entrepreneurs are rewarded with profit precisely to compensate them for personally taking on the uninsurable risks of running a business. Indeed, their willingness to bear those uninsurable risks means that the providers of the firm's other resource inputs (of land, labor, and capital) can almost completely ignore those risks.

To see why this is true, again consider our start-up company, Upside, and suppose that its two entrepreneurs used $100,000 of their own money to fund the firm. If revenues ever run below costs, the entrepreneurs will use that pile of money to cover the losses and make sure that the firm's workers and other resource suppliers get paid on time and in full. The resource suppliers are shielded from the losses because the entrepreneurs have taken it upon themselves to bear the firm's uninsurable financial risks. In such situations, no bill is sent to the workers or other resource suppliers asking them to help make up the firm's losses. The entrepreneurs, who took on the firm's uninsurable risks, are "on the hook."

The entrepreneurs' compensation for providing entrepreneurship and shielding the other resource suppliers from uninsurable risk is the opportunity to run the business and keep the firm's profits if things go well. Thus, entrepreneurship boils down to a simple bargain: In exchange for making sure that everyone else gets paid if things go wrong, the entrepreneur gets to receive the firm's profits if things go right.

Sources of Economic Profit

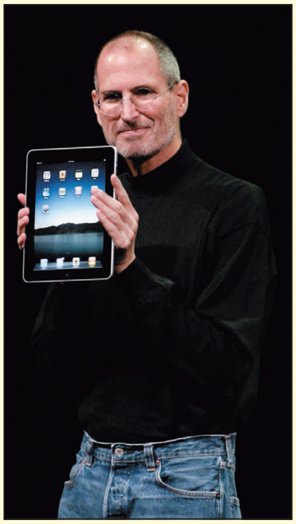
For bearing a firm's uninsurable risks, its entrepreneurs receive control of the firm. This allows them to try to make as much profit as possible in exchange for taking on those risks.

Along those lines, there are three main ways in which entrepreneurs can generate economic profits (that is, accounting profits that exceed normal profits):

- ***Create popular new products*** If an entrepreneur can develop a popular new product at a sufficiently low cost, his firm will be able to generate economic profits until competitors bring out competing products.

- ***Reduce production costs below rivals' costs*** Entrepreneurs who implement more-efficient production methods for existing products can generate economic profits until their efficiency gains are matched or exceeded by competitors.
- ***Create and maintain a profitable monopoly*** Entrepreneurs who possess a monopoly for their product may be able to generate economic rent by restricting their outputs and raising their prices. And such economic profits may persist if entry to the industry is blocked. But remember from Chapter 12 that having a monopoly does not guarantee that a monopoly will be profitable. If demand is weak relative to production costs, monopolies can and will go bankrupt.

CONSIDER THIS . . .



Apple CEO Steve Jobs

Before his death in 2011, entrepreneur Steven Jobs made billions of dollars in economic profit at the two major companies he headed, Apple Computers and Pixar Animation Studios.

He made these economic profits by creating wildly popular versions of existing products. Pixar, for example, was not the first animation studio. But it was the first to develop popular digitally animated movies, starting with *Toy Story* in 1995.

In the same way, Apple did not invent the personal computer, the MP3 music player, the Internet-capable mobile phone, or the touch-screen computer. But its iMac personal computer, iPod music player, Internet-capable iPhone, and iPad touch-screen computer were all major hits generating billions of dollars of profits.

Jobs achieved these successes largely because he defied conventional wisdom. His key decision was that Apple should develop both its own software as well as its own hardware. By contrast, computer-making rivals like Dell and IBM only made hardware while leaving the production of software up to other firms such as Microsoft. By building both its own hardware *and* its own software, Apple could create products in which the hardware and software were fully integrated and worked nearly flawlessly together.

The result was a string of hit products that rival firms could not easily duplicate because they were in nearly all cases either hardware makers *or* software makers but not both. Thus, another source of Apple's large profits has been a mild amount of monopoly power resulting from the inability of rivals to easily mimic Apple's strategy of producing both software and hardware.

By reallocating resources toward the production of popular new products that consumers prefer to old products, entrepreneurs improve allocative efficiency. By reducing production costs, they improve productive efficiency. Thus, with the important exception of monopoly, the entrepreneur's pursuit of profit clearly benefits society.

With monopolies, though, things are very different, because a monopoly's profit cannot be justified as rewarding its entrepreneurs for personally taking on a firm's financial risks or for making perceptive business decisions that enhance productive or allocative efficiency. As a result, governments use antitrust laws to impede monopolies from forming or, if they arise, break them up or restrict their business behaviors. In some instances, governments also regulate away their excessive profits.

Profit Rations Entrepreneurship

Even if a firm is making an accounting profit, its entrepreneurs may decide to quit and go into another line of business. This is because they will want their accounting profit to be at least as large as the typical accounting profit that their risk taking and decision making could on average be expected to earn in other business ventures. This typical accounting profit is known as their **normal profit.**

Because entrepreneurs will be comparing their current accounting profit with the normal profit that they could be making elsewhere, profit can be thought of as the "price" that allocates the scarce resource of entrepreneurship toward different possible economic activities. Just as wages must be paid to attract and retain labor, profits must be paid to attract and retain entrepreneurial talent. And just as higher wages attract workers, higher profits attract entrepreneurs.

Entrepreneurs, Profits, and Corporate Stockholders

In the actual economy, economic profits are distributed widely beyond the entrepreneurs who first start new businesses. The corporate structure of business enterprise has allowed millions of individuals to purchase ownership

LAST WORD

Determining the Price of Credit

A Variety of Lending Practices May Cause the Effective Interest Rate to Be Quite Different from What It Appears to Be.

Borrowing and lending—receiving and granting credit—are a way of life. Individuals receive credit when they negotiate a mortgage loan and when they use their credit cards. Individuals make loans when they open a savings account in a commercial bank or buy a government bond.

It is sometimes difficult to determine exactly how much interest we pay and receive when we borrow and lend. Let's suppose that you borrow \$10,000 that you agree to repay plus \$1,000 of interest at the end of 1 year. In this instance, the interest rate is 10 percent per year. To determine the interest rate i, we compare the interest paid with the amount borrowed:

$$i = \frac{\$1{,}000}{\$10{,}000} = 10\%$$

But in some cases a lender—say, a bank—will discount the interest payment from the loan amount at the time the loan is made. Thus, instead of giving the borrower \$10,000, the bank discounts the \$1,000 interest payment in advance, giving the borrower only \$9,000. This increases the interest rate:

$$i = \frac{\$1{,}000}{\$9{,}000} = 11\%$$

While the absolute amount of interest paid is the same, in the second case the borrower has only \$9,000 available for the year.

An even more subtle point is that, to simplify their calculations, many financial institutions assume a 360-day year (twelve 30-day months). This means the borrower has the use of the lender's funds for 5 days less than the normal year. This use of a

shares in corporations and therefore to share in the risks and rewards of ownership.

Some of these people participate in the for-profit economy by purchasing the stock of individual firms or by investing in mutual funds, which in turn buy corporate stock. Millions of additional people share in the profits of corporations through the financial investments of their pension funds. But at their core, all of the profits that are shared with these direct and indirect corporate shareholders are made possible in the first place by the activities of entrepreneurs. For instance, without Bill Gates and Paul Allen, there would have been no Microsoft Corporation—and, consequently, no Microsoft profits (dividends) to distribute to the millions of owners of Microsoft stock.

The desire on the part of so many millions of people to own corporate stock and thereby share in the risks and rewards of business reflects the fact that economic profit is the main energizer of the capitalistic economy. It influences both the level of economic output and the allocation of resources among alternative uses. The expectation of economic profit motivates firms to innovate. Innovation stimulates new investment, thereby increasing total output and employment. Thus, the pursuit of profit enhances economic growth by promoting innovation.

Profit also helps allocate resources among alternative lines of production, distribution, and sales. Entrepreneurs seek profit and shun losses. The occurrence of continuing profits in a firm or industry is a signal that society wants that particular firm or industry to expand. It attracts resources from firms and industries that are not profitable. But the rewards of profits are more than an inducement for a firm to expand; they also attract "short year" also increases the actual interest rate paid by the borrower.

The interest rate paid may change dramatically if a loan is repaid in installments. Suppose a bank lends you $10,000 and charges interest in the amount of $1,000 to be paid at the end of the year. But the loan contract requires that you repay the $10,000 loan in 12 equal monthly installments. As a result, the average amount of the loan outstanding during the year is only $5,000. Therefore:

$$i = \frac{\$1{,}000}{\$5{,}000} = 20\%$$

Here interest is paid on the total amount of the loan ($10,000) rather than on the outstanding balance (which averages $5,000 for the year), making for a much higher interest rate.

Another factor that influences the effective interest rate is whether or not interest is compounded. Suppose you deposit $10,000 in a savings account that pays a 10 percent interest rate compounded semiannually. In other words, interest is paid twice a year. At the end of the first 6 months, $500 of interest (10 percent of $10,000 for half a year) is added to your account. At the end of the year, interest is calculated on $10,500 so that the second interest payment is $525 (10 percent of $10,500 for half a year). Thus:

$$i = \frac{\$1{,}025}{\$10{,}000} = 10.25\%$$

This 10.25 percent return means that a bank offering a 10 percent interest rate compounded semiannually would pay more interest to its customers than a competitor offering a simple (noncompounded) rate of, say, 10.2 percent.

Two pieces of legislation have attempted to clarify interest charges and payments. The Truth in Lending Act of 1968 requires that lenders state the costs and terms of consumer credit in concise and uniform language, in particular, as an annual percentage rate (APR). More recently, the Truth in Savings Act of 1991 requires that all advertisements of deposit accounts by banks and other financial institutions disclose all fees connected with such accounts and the interest rates and APRs on each account. Nevertheless, some "payday" check-cashing firms that lend money to people in return for postdated personal checks have been found to receive interest payments on one- and two-week loans that are equivalent to APRs of hundreds of percent per year. These interest rates prompted calls for state legislators to protect consumers from "predatory lenders."

More recently, many banks have established fee-based "bounce (overdraft) protection" for checking accounts. The bank agrees to pay each overdraft for a flat fee of around $35. These fees are essentially interest on a loan for the amount of the overdraft. When the overdraft amount is small, the annual interest on the loan can easily exceed 1,000 percent.

Similarly, late-payment fees on credit card accounts can boost the actual interest rate paid on credit card balances to extremely high levels. Furthermore, low "teaser" rates designed to attract new customers often contain "fine print" that raises the interest rate to 16 percent, or even 28 percent, if a payment on the account is late. Also, low initial rates on some variable rate mortgages eventually "reset" to higher rates, greatly increasing the monthly payments that are due. "Let the borrower (or depositor) beware" remains a fitting motto in the world of credit.

the financing needed for expansion. In contrast, continuing losses penalize firms or industries that fail to adjust their productive efforts to match consumer wants. Such losses signal society's desire for the afflicted entities to contract.

So, in terms of our chapter-opening question, Walmart garnered large profits because it was locating its stores close to customers and delivering the mix of products many consumers wanted at exceptionally low prices. These profits signaled that society wanted more of its scarce resources allocated to Walmart stores. MF Global, in contrast, was not delivering products equivalent in value to the costs of the resources used to provide them—so the firm suffered losses. The losses and MF Global's bankrupty signaled that society would benefit from a reallocation of all or a part of those resources to some other use.

QUICK REVIEW 16.3

- Pure or economic profit is what remains after all explicit and implicit costs (including a normal profit) are subtracted from a firm's total revenue.
- As residual claimants, entrepreneurs receive accounting profit. They earn economic profit if their accounting profit exceeds the normal profit they could earn as entrepreneurs elsewhere.
- Economic profit has three sources: the bearing of uninsurable risk, the uncertainty of innovation, and monopoly power.
- The corporate structure of business enables millions of individuals to share in the risks and rewards of enterprise.
- Profit and profit expectations affect the levels of investment, total spending, and domestic output; profit and loss also allocate resources among alternative uses.

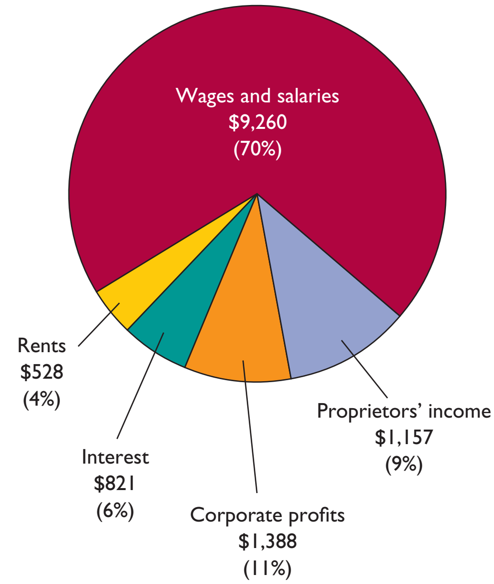
Income Shares

LO16.7 List the share of U.S. earnings received by each of the factors of production.

Our discussion in this and in the preceding chapter would not be complete without a brief examination of how U.S. income is distributed among wages, rent, interest, and profit.

Figure 16.3 shows how the income generated in the United States in 2011 was distributed among the five "functional" income categories tracked by the U.S. government. These five categories do not match up perfectly with the economic definitions of wages, rent, interest, and profit. The biggest difference is "proprietors' income," which is the income received by doctors, lawyers, small-business owners, farmers, and the owners of other unincorporated enterprises. In terms of our four economic categories, proprietors' income is a combination of wages and profit. The wages compensate for labor, while the profits compensate for entrepreneurship. The economists who have looked into this combination believe that the large majority of proprietors' income is implicitly composed of wages and salaries rather than profit. That is, the large majority of proprietors' income is compensation for labor rather than compensation for entrepreneurship.

FIGURE 16.3 The functional distribution of U.S. income, 2011. Seventy percent of U.S. income is received as wages and salaries. Income to property owners—corporate profit, interest, and rents—accounts for about 21 percent of total income. The dollar amounts are in billions of dollars.



Source: Bureau of Economic Analysis, **www.bea.gov**.

Taking that into account, note the dominant role of labor income in the U.S. economy. Even with labor income narrowly defined as wages and salaries, labor receives 70 percent of all income earned by Americans in a typical year. But if we add in proprietors' income because most of it is believed to be a payment for labor, then labor's share of national income rises to almost 80 percent, a percentage that has been remarkably stable since at least 1900. That leaves about 20 percent for "capitalists" in the form of rent, income, and profit. Ironically, income from capital is a relatively small share of the U.S. economy, though we call it a capitalist system.

SUMMARY

LO16.1 Explain the nature of economic rent and how it is determined.

Economic rent is the price paid for the use of land and other natural resources whose total supplies are fixed.

Differences in land rent result from differences in demand, often stemming from differences in the fertility and climate features of the land or differences in location.

Because the supply of land is fixed by nature, rent is a surplus payment that is socially unnecessary from the viewpoint of causing land to be supplied. The idea of land rent as a surplus payment gave rise to the single-tax movement of the late 1800s.

Although land rent is a surplus payment rather than a cost to the economy as a whole, to individuals and firms, land rents are correctly regarded as costs. The payment of land rents by individuals and firms is socially useful because it puts an opportunity cost on the use of land so that people are incentivized to put each piece of land to its best possible use.

LO16.2 Define interest and explain how interest rates vary based on risk, maturity, loan size, and taxability.

Interest is the price paid for the use of money. Because money is not itself an economic resource, people do not value money for its own sake; they value it for its purchasing power.

The interest rate on money loans determines the interest income earned by households for providing capital to firms.

Interest rates vary in size because loans differ as to risk, maturity, amount, and taxability.

The pure rate of interest is the hypothetical interest rate that would serve purely and solely to compensate lenders for their willingness to patiently forgo alternative consumption and investment opportunities until their money is repaid. The pure rate is best approximated by the interest rate on long-term, virtually riskless, 20-year U.S. Treasury bonds.

LO16.3 Explain the loanable funds theory of interest rates.

In the loanable funds theory of interest, the equilibrium interest rate in a loan market is determined by the demand for and supply of loanable funds in that market. Other things equal, an increase in the supply of loanable funds reduces the equilibrium interest rate, whereas a decrease in supply increases it. And increases in the demand for loanable funds raise the equilibrium interest rate, whereas decreases in demand reduce it.

LO16.4 Demonstrate how interest rates relate to the time-value of money.

The time-value of money is the idea that $1 today has more value than $1 sometime in the future because the $1 today can be placed in an interest-bearing account and earn compound interest over time. Future value is the amount to which a current amount of money will grow through interest compounding. Present value is the current value of some money payment to be received in the future.

LO16.5 Explain the role of interest rates in allocating capital, modulating R&D spending, and helping to determine the economy's total output of goods and services.

The equilibrium interest rate influences the level of investment and helps ration financial and physical capital to specific firms and industries. Similarly, this rate influences the size and composition of R&D spending. The real interest rate, not the nominal rate, is critical to investment and R&D decisions.

Although designed to make funds available to low-income borrowers, usury laws tend to allocate credit to high-income persons, subsidize high-income borrowers at the expense of lenders, and lessen the efficiency with which loanable funds are allocated.

LO16.6 Relate why economic profits occur, and how profits, along with losses, allocate resources among alternative uses.

As residual claimants, entrepreneurs receive a firm's accounting profits (total revenue minus explicit costs) in exchange for assuming the uninsurable risks associated with running a business. An entrepreneur can earn an economic profit (total revenue minus both explicit and implicit costs, including a normal profit) if her firm's accounting profit exceeds the normal profit that her entrepreneurship could on average earn in other business ventures.

The corporate form of business organization has allowed the millions who own corporate stock to share in the financial risks and economic profits engendered by entrepreneurship. Profits are the key energizer of business firms within the capitalist system. Profit expectations influence innovating and investment activities and therefore the economy's levels of employment and economic growth. The basic function of profits and losses, however, is to allocate resources in accord with consumers' preferences.

LO16.7 List the share of U.S. earnings received by each of the factors of production.

The largest share of all income earned by Americans—about 70 percent—goes to labor, a share narrowly defined as "wages and salaries." When labor's share is more broadly defined to include "proprietors' income," it rises to about 80 percent of national income, leaving about 20 percent for rent, interest, and profit payments to the providers of land, capital, and entrepreneurship.

TERMS AND CONCEPTS

economic rent	compound interest	explicit costs
incentive function	future value	implicit costs
single-tax movement	present value	economic or pure profit
pure rate of interest	nominal interest rate	insurable risks
loanable funds theory of interest	real interest rate	uninsurable risks
time-value of money	usury laws	normal profit

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. How does the economist's use of the term "rent" differ from everyday usage? Explain: "Though rent need not be paid by society to make land available, rental payments are useful in guiding land into the most productive uses." **LO16.1**

2. Explain why economic rent is a surplus payment when viewed by the economy as a whole but a cost of production from the standpoint of individual firms and industries.

Explain: "Land rent performs no 'incentive function' for the overall economy." LO16.1

3. How does Henry George's proposal for a single tax on land relate to the elasticity of the supply of land? Why are there so few remaining advocates of George's proposal? LO16.1
4. If money is not an economic resource, why is interest paid and received for its use? What considerations account for the fact that interest rates differ greatly on various types of loans? Use those considerations to explain the relative sizes of the interest rates on the following: LO16.2
   - **a.** A 10-year $1,000 government bond.
   - **b.** A $20 pawnshop loan.
   - **c.** A 30-year mortgage loan on a $175,000 house.
   - **d.** A 24-month $12,000 commercial bank loan to finance the purchase of an automobile.
   - **e.** A 60-day $100 loan from a personal finance company.
5. Why is the supply of loanable funds upsloping? Why is the demand for loanable funds downsloping? Explain the equilibrium interest rate. List some factors that might cause it to change. LO16.3
6. Here is the deal: You can pay your college tuition at the beginning of the academic year or the same amount at the end of the academic year. You either already have the money in an interest-bearing account or will have to borrow it. Deal, or no deal? Explain your financial reasoning. Relate your answer to the time-value of money, present value, and future value. LO16.4
7. What are the major economic functions of the interest rate? How might the fact that many businesses finance their investment activities internally affect the efficiency with which the interest rate performs its functions? LO16.5
8. Distinguish between nominal and real interest rates. Which is more relevant in making investment and R&D decisions? If the nominal interest rate is 12 percent and the inflation rate is 8 percent, what is the real rate of interest? LO16.5
9. Historically, usury laws that put below-equilibrium ceilings on interest rates have been used by some states to make credit available to poor people who could not otherwise afford to borrow. Critics contend that poor people are those most likely to be hurt by such laws. Which view is correct? LO16.5
10. How do the concepts of accounting profit and economic profit differ? Why is economic profit smaller than accounting profit? What are the three basic sources of economic profit? Classify each of the following according to those sources: LO16.6
    - **a.** A firm's profit from developing and patenting a new medication that greatly reduces cholesterol and thus diminishes the likelihood of heart disease and stroke.
    - **b.** A restaurant's profit that results from the completion of a new highway past its door.
    - **c.** The profit received by a firm due to an unanticipated change in consumer tastes.
11. Why is the distinction between insurable and uninsurable risks significant for the theory of profit? Carefully evaluate: "All economic profit can be traced to either uncertainty or the desire to avoid it." What are the major functions of economic profit? LO16.6
12. What is the combined rent, interest, and profit share of the income earned by Americans in a typical year if proprietors' income is included within the labor (wage) share? LO16.7
13. **LAST WORD** Assume that you borrow $5,000, and you pay back the $5,000 plus $250 in interest at the end of the year. Assuming no inflation, what is the real interest rate? What would the interest rate be if the $250 of interest had been discounted at the time the loan was made? What would the interest rate be if you were required to repay the loan in 12 equal monthly installments?

REVIEW QUESTIONS

1. When using a supply-and-demand model to illustrate how land rents are set, economists typically draw the supply curve as a vertical line because: LO16.1
   - **a.** The supply of land is fixed.
   - **b.** The supply of land is perfectly inelastic.
   - **c.** The quantity supplied of land does not increase when rents go up.
   - **d.** All of the above.
2. In the 1980s land prices in Japan surged upward in a "speculative bubble." Land prices then fell for 11 straight years between 1990 and 2001. What can we safely assume happened to *land rent* in Japan over those 11 years? Use graphical analysis to illustrate your answer. LO16.1
3. The a main argument put forth by advocates of the *single-tax movement* was that: LO16.1
   - **a.** Taxing only income would make for a more equal society.
   - **b.** Taxing only land would be very efficient because taxing land does not decrease its supply.
   - **c.** Taxing only imports would help to protect local jobs and stimulate local entrepreneurs.
   - **d.** Having only one tax would be much easier for people to understand and much less costly to administer than our current system with its wide variety of taxes.
4. Angela puts $1,000 in a savings account that pays 3 percent per year. What is the future value of her money one year from now? LO16.4
   - **a.** $970.
   - **b.** $1,000.
   - **c.** $1,003.
   - **d.** $1,030.
5. As shown in Table 16.2, $1,000 invested at 10 percent compound interest will grow into $1,331 after three years. What

is the present value of $2,662 in three years if it is discounted back to the present at a 10 percent compound interest rate? (Hint: $2,662 is twice as much as $1,331.) **LO16.4**

6. Entrepreneurs are the residual claimants at their respective firms. This means that they: **LO16.6**
   a. Only get paid if there is any money left over after all the other factors of production have been paid.
   b. Must bear the financial risks of running their firms.
   c. Receive whatever accounting profits or losses their firms generate.
   d. All of the above.
7. True or false. As a capitalist economy, the vast majority of U.S. national income flows to the owners of capital. **LO16.7**

PROBLEMS

1. Suppose that you own a 10-acre plot of land that you would like to rent out to wheat farmers. For them, bringing in a harvest involves $30 per acre for seed, $80 per acre for fertilizer, and $70 per acre for equipment rentals and labor. With these inputs, the land will yield 40 bushels of wheat per acre. If the price at which wheat can be sold is $5 per bushel and if farmers want to earn a normal profit of $10 per acre, what is the most that any farmer would pay to rent your 10 acres? What if the price of wheat rose to $6 per bushel? **LO16.1**
2. Suppose that the demand for loanable funds for car loans in the Milwaukee area is $10 million per month at an interest rate of 10 percent per year, $11 million at an interest rate of 9 percent per year, $12 million at an interest rate of 8 percent per year, and so on. If the supply of loanable funds is fixed at $15 million, what will be the equilibrium interest rate? If the government imposes a usury law and says that car loans cannot exceed 3 percent per year, how big will the monthly shortage (or excess demand) for car loans be? What if the usury limit is raised to 7 percent per year? **LO16.3**
3. To fund its wars against Napoleon, the British government sold consol bonds. They were referred to as "perpetuities" because they would pay £3 every year in perpetuity (forever). If a citizen could purchase a consol for £25, what would its annual interest rate be? What if the price were £50? £100? Bonds are known as "fixed income" securities because the future payments that they will make to investors are fixed by the bond agreement in advance. Do the interest rates of bonds and other investments that offer fixed future payments vary positively or inversely with their current prices? **LO16.4**
4. Suppose that the interest rate is 4 percent. What is the future value of $100 four years from now? How much of the future value is total interest? By how much would total interest be greater at a 6 percent interest rate than at a 4 percent interest rate? **LO16.4**
5. You are currently a worker earning $60,000 per year but are considering becoming an entrepreneur. You will not switch unless you earn an accounting profit that is on average at least as great as your current salary. You look into opening a small grocery store. Suppose that the store has annual costs of $150,000 for labor, $40,000 for rent, and $30,000 for equipment. There is a one-half probability that revenues will be $200,000 and a one-half probability that revenues will be $400,000. **LO16.6**
   a. In the low-revenue situation, what will your accounting profit or loss be? In the high-revenue situation?
   b. *On average*, how much do you expect your revenue to be? Your accounting profit? Your economic profit? Will you quit your job and try your hand at being an entrepreneur?
   c. Suppose the government imposes a 25 percent tax on accounting profits. This tax is only levied if a firm is earning positive accounting profits. What will your after-tax accounting profit be in the low-revenue case? In the high-revenue case? What will your *average* after-tax accounting profit be? What about your *average* after-tax economic profit? Will you now want to quit your job and try your hand at being an entrepreneur?
   d. Other things equal, does the imposition of the 25 percent profit tax increase or decrease the supply of entrepreneurship in the economy?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **17**

Natural Resource and Energy Economics

Learning Objectives

LO17.1 Explain why falling birthrates mean that we are not likely to run out of natural resources.

LO17.2 Describe why using a mix of energy sources is efficient, even if some of them are quite costly.

LO17.3 Discuss why running out of oil would not mean running out of energy.

LO17.4 Show how the profit motive can encourage resource conservation.

LO17.5 Relate how to use property rights to prevent deforestation and species extinction.

People like to consume goods and services. But to produce those goods and services, natural resources must be used up. Some natural resources, such as solar energy, forests, and fish, are renewable and can potentially be exploited indefinitely. Other resources, such as oil and coal, are in fixed supply and can be used only once. This chapter explores two issues in relation to our supplies of resources and energy. The first is whether we are likely to run out of resources in the near or even distant future and thereby face the possibility of either a drastic reduction in living standards or even, perhaps, the collapse of civilization as we know it. The second is how to best utilize and manage our resources so that we can maximize the benefits that we receive from them both now and in the future.

We begin the chapter by addressing the issue of whether we are about to run out of resources. We then turn to energy economics and natural resource economics, focusing on the incentive structures that help to promote conservation and sustainability.

Resource Supplies: Doom or Boom?

LO17.1 Explain why falling birthrates mean that we are not likely to run out of natural resources.

Since the beginning of the Industrial Revolution in the late 18th century, a historically unprecedented increase in both population and living standards has taken place. The world's population has increased from 1 billion people in 1800 to about 7 billion today, and the average person living in the United States enjoys a standard of living at least 12 times higher than that of the average American living in 1800. Stated slightly differently, many more people are alive today and levels of consumption per person are much higher. These two factors mean that human beings are now consuming vastly more resources than before the Industrial Revolution both in absolute terms and in per capita terms. This fact has led many observers to wonder if our current economic system and its high living standards are sustainable. In particular, will the availability of natural resources be sufficient to meet the growing demand for them?

A sensible response clearly involves looking at *both* resource demand and resource supply. We begin by examining human population growth because larger populations mean greater demand for resources.

Population Growth

We can trace the debate over the sustainability of resources back to 1798, when an Anglican minister in England named Thomas Malthus published *An Essay on the Principle of Population*. In that essay, Malthus argued that human living standards could only temporarily rise above subsistence levels. Any temporary increase in living standards would cause people to have more children and thereby increase the population. With so many more people to feed, per capita living standards would be driven back down to subsistence levels.

Falling Birthrates Unfortunately for Malthus's theory—but fortunately for society—higher living standards have *not* produced higher birthrates. In fact, just the opposite has happened. Higher standards of living are associated with *lower* birthrates. Birthrates are falling rapidly throughout the world and the majority of the world's population is now living in countries that have birthrates that are lower than the **replacement rate** necessary to keep their respective populations stable over time.

Table 17.1 lists the total fertility rates for 12 selected nations including the United States. The **total fertility rate** is the average number of children that a woman is expected to have during her lifetime. Taking into account infant and child mortality, a total fertility rate of about 2.1 births per woman per lifetime is necessary to keep the population constant, since 2.1 children equals 1 child to replace the mother, 1 child to replace the father, and 0.1 extra child who can be expected to die before becoming old enough to reproduce.

TABLE 17.1 **Total Fertility Rates for Selected Countries, 2012**

Country	Total Fertility Rate
Australia	1.77
Canada	1.59
China	1.55
France	2.08
Germany	1.41
Hong Kong	1.09
Italy	1.40
Japan	1.39
Russia	1.61
South Korea	1.23
Sweden	1.67
United States	2.06

Source: The World Factbook, **www.cia.gov**. Data are 2012 estimates.

As you can see from Table 17.1, total fertility rates in many nations are well below the 2.1 rate necessary to keep the population stable over time. As a result, populations are expected to fall rapidly in many countries over the next few decades, with, for instance, the population of Russia expected to fall by about 12 percent from its current level of 142 million people to fewer than 125 million in 2050. And Russia is not alone; 28 countries are expected to see their populations fall by at least 10 percent by 2050, and of these, 17 are expected to experience a decline of at least 20 percent by 2050.

Worldwide, the precipitous fall of birthrates means that many **demographers** (scientists who study human populations) now expect the world's population to reach a peak of 9 billion people or fewer sometime around the middle of this century before beginning to fall, perhaps quite rapidly. For instance, if the worldwide total fertility rate declines to 1 birth per woman per lifetime (which is near Hong Kong's current rate of 1.09 per woman per lifetime), then each generation will be only half as large as the previous one because there will be only one child on average for every two parents. And even a rate of 1.3 births per woman per lifetime will reduce a country's population by half in just under 45 years.

The Demographic Transition The world's population increased so rapidly from 1800 to the present day because the higher living standards that arrive when a country begins to modernize bring with them much lower death rates. Before modernization happens, death rates are typically so high that women have to give birth to more than six children per lifetime just to ensure that, on average, two will survive to adulthood. But once living standards begin to rise and modern medical care becomes available, death rates plummet so that nearly all children survive to adulthood. This causes a temporary population explosion because parents—initially unaware that such a revolutionary change in death rates has taken place—for a while keep having six or more children. The impression persists that they must have many children to ensure that at least two will survive to adulthood. The result is one or two generations of very rapid population growth until parents adjust to the new situation and reduce the number of children that they choose to have. Demographers refer to this three-step shift from (1) the traditional situation of simultaneously high birth and death rates rates through (2) a transition period of high birth rates and low death rates and then finally on to (3) the current situation of simultaneously low birth and death rates as the **demographic transition.**

The overall world population is still increasing because many countries such as India and Indonesia began modernizing only relatively recently and are still in the transition phase where death rates have fallen but birthrates are still relatively high. Nevertheless, birthrates are falling rapidly nearly everywhere. This means that the end of rapid population growth is at hand. Furthermore, because fertility rates tend to fall below the replacement rate as countries modernize, we can also expect total world population to begin to decline during the twenty-first century. This is a critical fact to keep in mind when considering whether we are likely to ever face a resource crisis: Fewer people means fewer demands placed on society's scarce resources.

CONSIDER THIS . . .

Can Governments Raise Birthrates?

Low birthrates pose major problems for countries encountering very slow population growth or actual population declines. The primary problem is that with very few children being born today, very few workers will be alive in a few decades to pay the large amounts of taxes that will be needed if governments are to keep their current promises regarding Social Security and other old-age pension programs. Too few young workers will be supporting too many elderly retirees. Another potential problem is a lack of soldiers. Consider Russia. With its population expected to fall by 12 percent by midcentury, defending its borders will be much harder.

As a response, Russian President Vladimir Putin announced a new policy in 2006 that would pay any Russian woman who chooses to have a second child a bounty worth 250,000 rubles ($9,280). In addition, the Russian government promised to double monthly child benefits in an effort to make having children less financially burdensome for parents. Many other countries have experimented with similar policies. In 2004, France began offering its mothers a payment of €800 ($1,040) for each child born and Italy began offering a €1,000 ($1,300) payment for second children.

As far as demographers can tell, however, these and other policies aimed at raising birthrates by offering maternity leave, free day care, or other subsidies to mothers or their children have not been able to generate any substantial increases in fertility levels in any country in which they have been attempted. For example, Australia's policy of paying mothers A$5,000 ($4,500) for each child increased the number of births by less than 4 percent. It therefore appears that the incentives to have more babies provided by these government plans are being swamped by the broader social and economic forces that are leading to declining overall fertility rates.

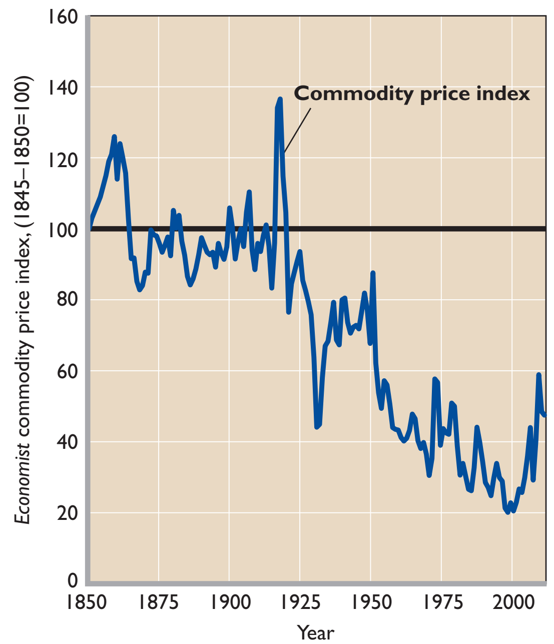
Possible Explanations for Low Fertility Demographers have been surprised at just how low fertility rates have fallen and why they have fallen so far below the replacement rate in so many countries. The decline of fertility rates to such low levels is especially surprising given the fact that couples typically tell demographers that they would like to have *at least* two children. Because this implies that most couples would prefer higher total fertility rates than we actually observe, it seems probable that social or economic factors are constraining couples to have fewer children than they desire, thereby causing total fertility rates to fall so low. Demographers have not yet reached agreement on which factors are most important, but possible candidates include changing attitudes toward religion, the much wider career opportunities available to women in modern economies, and the expense of raising children in modern societies. Indeed, children have been transformed from economic assets that could be put to work at an early age in agricultural societies into economic liabilities that are very costly to raise in modern societies where child labor is illegal and where children must attend school until adulthood. The nearby Consider This vignette discusses current government efforts to raise birthrates by offering financial incentives to parents.

Resource Consumption per Person

Thomas Malthus's tradition of predicting a collapse in living standards has been carried on to this day by various individuals and groups. One well-reported prediction was made by Stanford University butterfly expert Paul Ehrlich. In his 1968 book *The Population Bomb*, he made the Malthusian prediction that the population would soon outstrip resources so that "in the 1970s and 1980s hundreds of millions of people will starve to death in spite of any crash programs embarked upon now." Contrary to this prediction, no famines approaching these magnitudes materialized then and none appear likely today.

Falling Resource Prices One reason that Ehrlich's pessimism was not borne out was because the population growth rate slowed dramatically as living standards around the world rose. Another reason is that the long-run evidence indicates that the supply of productive resources available to be made into goods and services has been increasing faster than the demand for those resources for at least 150 years. This is best seen by looking at Figure 17.1, which tracks *The Economist* magazine's commodity price index for the years 1850 to 2011. The index currently contains 25 important commodities including aluminum, copper, corn, rice, wheat, coffee, rubber, sugar, and soybeans. In earlier days, it included commodities such as candle wax, silk, and indigo, which were important at the time. The index also adjusts for inflation so that one can see how the real cost of commodities has evolved over time and it is standardized so that the real price of commodities during the years 1845 to 1850 is given a value of 100.

FIGURE 17.1 ***The Economist's* commodity price index, 1850–2011.** *The Economist* magazine's commodity price index attempts to keep track of the prices of the commodities most common in international trade. It is adjusted for inflation and scaled so that commodity prices in the years 1845–1850 are set to an index value of 100. The figure shows that real commodity prices are volatile (vary considerably from year to year) but are now 60 percent lower than they were in the mid-nineteenth century. This implies that commodity supplies have increased faster than commodity demands.



Source: The Economist, **www.economist.com**. © The Economist Newspaper Limited, London (2012). Inflation adjustments made using the GDP deflator for the United States calculated by the Bureau of Economic Analysis, **www.bea.gov**.

As Figure 17.1 demonstrates, a dramatic long-run decline in real commodity prices has occurred. With the current value of the index at about 50, the real cost of buying commodities today is roughly 50 percent lower than it was in the initial 1845–1850 period. This means that commodity supplies have increased faster than commodity demands, since the only way that commodity prices could have fallen so much in the face of increasing demand is if the supply curve for commodities shifted to the right faster than the demand curve for commodities shifted to the right.

A key point is that the long-run fall of commodity prices implies that commodity supplies have grown faster than the sum total of the two pressures that have acted over this time to increase commodity demands. The first is the huge rise in the total number of people alive and therefore consuming resources (since 1850, the world's population has risen from 1.25 billion to 7.1 billion). The second is the huge rise in the amount of consumption *per person*. That is, more people are alive today than in 1850, and each person alive today is on average consuming several times more than was consumed by the average person alive in 1850. Still, the long-run fall in commodity prices confirms that supplies have managed to grow fast enough to overcome both these demand-increasing pressures.

But will supplies be able to overcome these two pressures in the future? Prospects are hopeful. First, the rapid and continuing decline in birthrates means that the huge population increases that occurred during the nineteenth and twentieth centuries are not likely to continue in the future. Indeed, we have seen that population decline has begun in several countries and it now seems likely that overall world population will begin to decline within this century. This trend will moderate future increases in the total demand for goods and services. Second, resource consumption *per person* (as distinct from goods and services consumption per person) also has either leveled off or declined in the past decade or so in the richest countries, which currently consume the largest fraction of the world's resources.

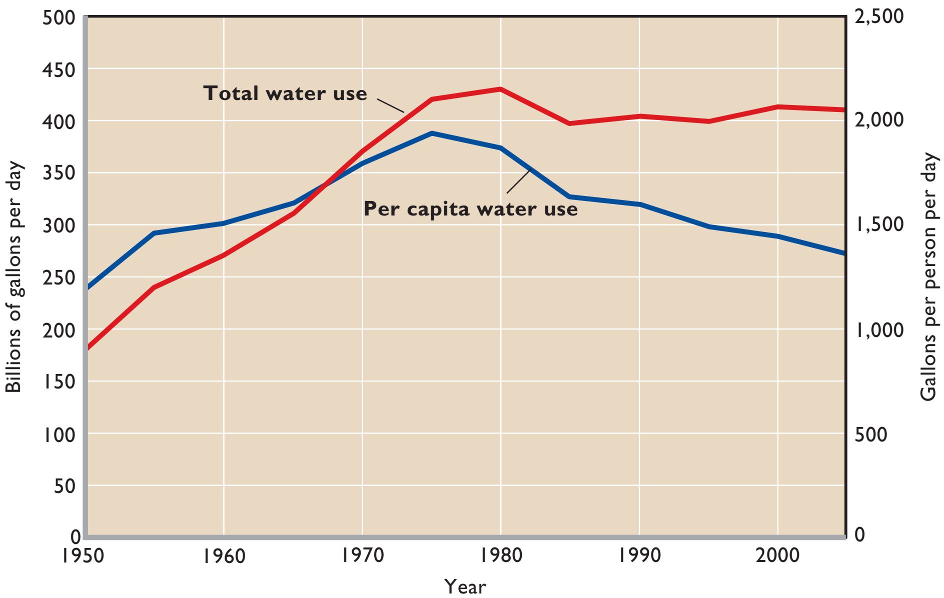


FIGURE 17.2 Total and per capita water use in the United States, 1950–2005. Average total water use in the United States peaked at 440 billion gallons per day in 1980 before declining to about 400 billion gallons per day in 1985, where it remained through 2005, the last year for which data were available. Average per capita water consumption fell 28 percent from a peak of 1,941 gallons per person per day in 1975 to only 1,363 gallons per person per day in 2005. These data are reported every 5 years. The 2010 data were not available at the time of publication, but they may be available when you read this. If interested, check the U.S. Geological Survey Web site.

Source: Estimated Use of Water in the United States in 2005, United States Geological Survey, **www.usgs.gov**.

Consumption Trends for Water, Energy, and Materials The leveling off or decline of per capita resource consumption can be observed in Figures 17.2, 17.3, and 17.4, which show, respectively, how much water, energy, and other resources have been consumed on a daily or annual basis in both total and per capita terms over the last few decades in the United States. The red lines in each figure show total use while the blue lines trace per capita use. To accommodate both sets of data, the units measuring total use are on the vertical scales on the left side of each figure while the units measuring per capita use are shown on the vertical scales on the right side of each figure.

The blue line in Figure 17.2 shows that per capita water use in the United States peaked in 1975 at 1,941 gallons per person per day. It then fell by 30 percent to just 1,363 gallons per person per day in 2005. The blue line in Figure 17.3 shows that annual per capita energy use peaked at 360 million British thermal units per person in 1979 before falling sharply during the early 1980s recession, gradually rising through the late 1990s and then falling again through 2011, when it ran at a rate of 312 BTUs per person per year. (A **British thermal unit,** or BTU, is the amount of energy required to raise the temperature of 1 pound of water by 1 degree Fahrenheit.)

Finally, Figure 17.4 takes advantage of a fundamental principle of physics to show that the per capita use of other resources has also leveled off since 1990. This principle

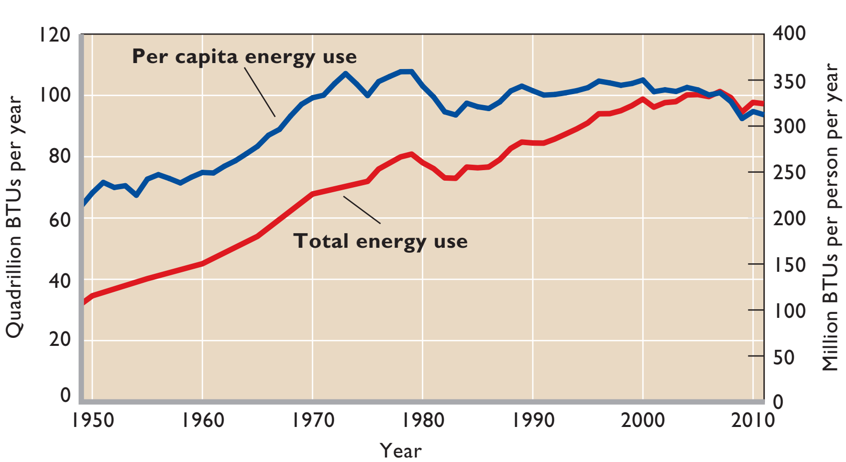
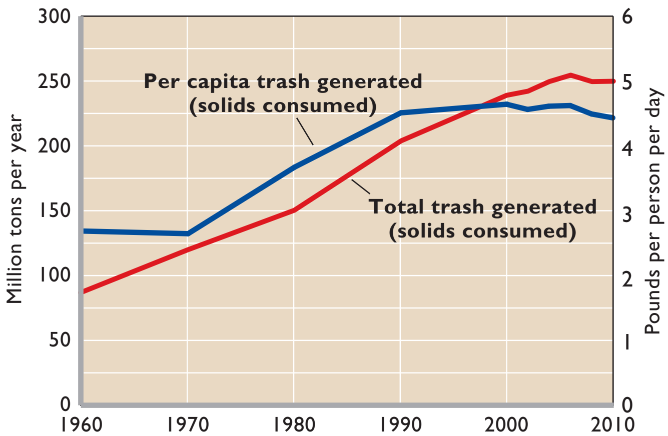


FIGURE 17.3 Total and per capita energy consumption in the United States, 1950–2011. Per capita energy consumption in the United States peaked at 360 million British thermal units (BTUs) per person per day in 1979. It fell dramatically during the early 1980s recession and then rose again until 1999, after which it has been mostly falling, down to a value of 312 million BTUs per person per day in 2011. Total energy consumption between 1950 and 2011 nearly tripled, increasing from 34.6 quadrillion BTUs in 1950 to 97.3 quadrillion BTUs in 2011. Since 1990, total energy consumption has increased by an average of only 0.7 percent per year.

Source: Annual Energy Review 2011, United States Energy Information Administration, **www.eia.doe.gov**.



Source: United States Environmental Protection Agency, **www.epa.gov**.

FIGURE 17.4 Total and per capita trash generation in the United States, 1960–2010. Although the total level of trash generated in the United States increased from 88.1 million tons in 1960 to 249.9 million tons in 2010, the amount of trash generated per person has held steady at approximately 4.5 pounds per person per day since 1990. Because *trash generated equals solids consumed,* we know that per capita consumption of solids has stayed relatively constant over the past 20 years.

states that matter is neither created nor destroyed—only transformed—by the sorts of chemical reactions that take place as raw materials are turned into finished products and then consumed. As a result, we can measure how much use of solid objects like plastics, metals, and paper takes place by measuring how much solid waste (commonly called trash or garbage) is generated when they are thrown away. Consequently, because Figure 17.4 shows that per capita trash generation has leveled off at about 4.5 pounds per person per day since 1990, we can conclude that per capita consumption of solids has also leveled off since that time.

The Likely Long-Run Decline in Resource Demands These three figures give further cause for optimism on the availability of future resource supplies. We have already provided evidence that the number of people in the world is not likely to increase substantially. Figures 17.2, 17.3, and 17.4 show that per capita consumption levels are also likely to either level off or decline. Together, these two facts suggest that the total demand for resources is likely to reach a peak in the relatively near future before falling over time as populations decline.

That being said, resource demand is likely to increase substantially for the next few decades as large parts of the world modernize and begin to consume as much per capita as the citizens of rich countries do today. For instance, per capita energy use in the United States in 2011 was 312 million BTUs per person. If every person in the world used that much energy, total annual energy demand would be 2,184 quadrillion BTUs, or about 4.5 times the 2009 world production of 484 quadrillion BTUs. One of the world's great economic challenges over the coming decades will be to supply the resources that will be demanded as living standards in poorer countries rise to rich-country levels. But because population growth rates are slowing and because per capita resource uses in rich countries have leveled off, we can now foresee a maximum total demand for resources even if living standards all over the world rise to rich-country levels. Given the ongoing improvements in technology and productivity that characterize modern economies and that allow us to produce increasingly more from any given set of inputs, it consequently seems unlikely that we will run into a situation where the total demand for resources exhausts their overall supply.

Regional Resource Challenges Significant challenges are likely to appear in those places where local supplies of certain resources are extremely limited. Water, for instance, is a rare and precious commodity in many places, including the Middle East and the American Southwest. Governments will have to work hard to ensure that the limited supplies of water in such areas are used efficiently and that disputes over water rights are settled peacefully. Along the same lines, resources are often produced in certain areas but consumed in others with, for instance, one-quarter of the world's oil being produced in the Middle East but most of the demand for oil coming from Europe, North America, and East Asia. In such cases, institutions must be developed that can move such resources from the areas in which they originate to the areas in which they are used. If not, local shortages may develop in the areas that cannot produce these resources despite the fact that the resources in question may at the same time be in very plentiful supply in the areas in which they are produced.

QUICK REVIEW 17.1

- Thomas Malthus and others have worried that increases in our demand for resources will outrun the supply of resources, but commodity prices have been falling for more than a century, indicating that supply has increased by more than demand.
- Because total fertility rates are very low and falling, population growth for the world will soon turn negative and thereby reduce the demand for natural resources.
- Per capita consumption of resources such as water, energy, and solids has either fallen or remained constant in the United States. If per capita consumption continues to stay the same or decrease while populations fall, total resource demand will fall—meaning that the demand for resources is unlikely to threaten to use up the available supply of resources.
- Significant increases in resource demands are likely over the next few decades, however, as living standards in poorer countries rise toward those in richer countries.

Energy Economics

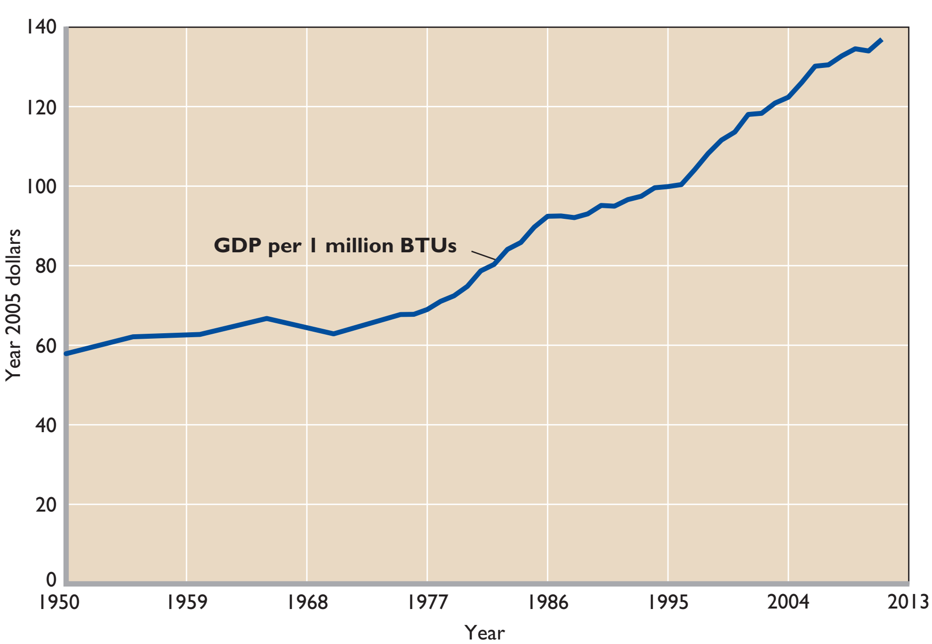
LO17.2 Describe why using a mix of energy sources is efficient, even if some of them are quite costly.

Energy economics studies how people deal with energy scarcity. This involves both demand and supply. In terms of energy supply, people are interested in attempting to find and exploit low-cost energy sources. But since energy is only one input into a production process, often the best energy source to use in a given situation is, paradoxically, actually rather expensive—yet still the best choice when other costs are taken into account. The economy therefore develops and exploits many different energy sources, from fossil fuels to nuclear power.

Energy Efficiency Is Increasing

In terms of energy demand, the most interesting fact is that per capita energy use has leveled off in recent years in developed countries, as we previously illustrated for the United States in Figure 17.3. This fact implies that our economy has become increasingly efficient at using energy to produce goods and services. This is best seen by noting that while per capita energy inputs fell from 338 BTUs per person per year in 1990 down to only 312 BTUs per person per year in 2011, real GDP per person rose during that time period by 39 percent. As a result, people were able to make and consume nearly two-fifths more goods and services per person despite using about 8 percent less energy per person.

This increase in energy efficiency has been part of a long historical trend, as Figure 17.5 makes quite clear. For the years 1950 through 2011, it shows the number of inflation-adjusted dollars of GDP that the U.S. economy has produced each year for every 1 million BTUs of energy consumed in the United States. The figure



Source: United States Energy Information Administration, **www.eia.doe.gov**.

FIGURE 17.5 Inflation-adjusted GDP per million BTUs of energy consumption in the United States, 1950–2011. This figure shows the number of dollars' worth of real GDP the U.S. economy produced per million BTUs of energy consumed in each year from 1950 through 2011 when annual GDP figures are converted to year 2005 dollars to account for inflation. Energy efficiency has more than doubled during this period, with real output per energy input rising from $57.90 worth of GDP per million BTUs in 1950 to $136.80 worth of GDP per million BTUs in 2011.

demonstrates that technological improvements greatly increased energy efficiency, so much so that although 1 million BTUs of energy yielded only $57.90 worth of goods and services in 1950, 1 million BTUs of energy yielded $136.80 worth of goods and services in 2011(when the comparison is made using year 2005 dollars to account for inflation).

Keep this huge increase in energy efficiency in mind when considering the magnitude of future energy demands. Because better technology means that more output can be produced with the same amount of energy input, rising living standards in the future will not necessarily depend on using more energy. The behavior of the U.S. economy since 1990 bears this out because, as we just pointed out, real GDP per person increased by nearly 40 percent between 1990 and 2011, while per capita energy inputs declined. Living standards can be raised without having to increase energy inputs.

Efficient Electricity Use

We just saw that the United States has grown increasingly efficient at using energy. The same is true for other developed countries. An interesting fact about energy efficiency, however, is that it often involves using a mix of energy inputs, some of which are much more expensive than others. The best way to see why this is true is to examine electric power generation.

The Challenge: Highly Variable Demand A typical electric plant has to serve tens of thousands of homes and businesses and is expected to deliver an uninterrupted supply of electricity 24 hours a day, 7 days a week. This task is not easy. The problem is that massive changes in energy demand occur over the course of a day. Demand is extremely low at night when people are sleeping, begins to rise rapidly in the morning as people wake up and turn on their lights, rises even more when they are at work, falls a bit as they commute home, rises back up a bit in the evening when they turn on their houselights to deal with the darkness and their televisions to deal with their boredom, and finally collapses as they turn out their lights and go to sleep.

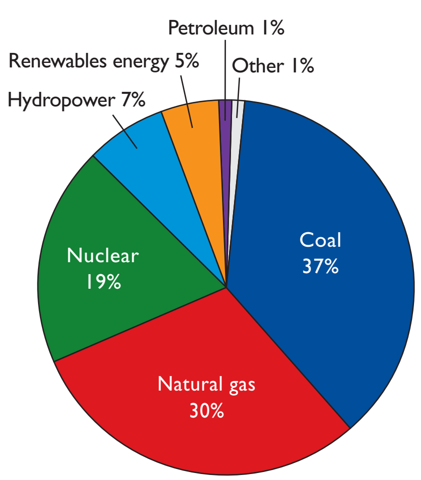
The problem for electric companies as they try to minimize the cost of providing for such large variations in the demand for electricity is that the power plants that have the lowest operating costs also have the highest fixed costs in terms of construction. For instance, large coal-fired plants can produce energy at a cost of about 4 cents per kilowatt hour. But they can do this only if they are built large enough to exploit economies of scale and if they are then operated at full capacity. To see why this can be a problem, imagine that such a plant has a maximum generating capacity of 20 megawatts per hour but that its customers' peak afternoon demand for electricity is 25 megawatts per hour. One solution would be to build two 20-megawatt coal-fired plants. But that would be very wasteful because one would be operating at full capacity (and hence minimum cost), while the other would be producing only 5 megawatts of its 20-megawatt capacity. Given that such plants cost hundreds of millions of dollars to build, this would be highly inefficient.

The Solution: Mixing Generation Technologies

The solution that electric companies employ is to use a mix of different types of generation technology. This turns out to be optimal because even though some electricity generation plants have very high operating costs, they have low fixed costs (that is, they are very inexpensive to build). Thus, the power company in our example might build one large coal-fired plant to generate 20 of the required 25 megawatts of energy at 4 cents per kilowatt hour, but it would then build a small 5-megawatt natural gas generator to supply the rest. Such plants produce electricity at the much higher cost of 15 cents per kilowatt hour, but they are relatively inexpensive to build. As a result, this solution would save the electric company from having to build a second very expensive coal-fired plant that would wastefully operate well below its full capacity.

The result of this process of mixing generator technologies is that the United States currently generates electricity from a variety of energy sources, as we show in Figure 17.6.

FIGURE 17.6 Percentages of U.S. electricity generated using various energy sources, 2012. About 37 percent of U.S. electricity was generated by coal-fired plants in 2012, with natural gas and nuclear power accounting together for a further 49 percent of the total.



Source: United States Energy Information Administration, **www.eia.doe.gov**.

About two-thirds of the total is generated at coal and gas plants while the rest comes from a variety of sources including hydroelectric power, natural gas, and renewable energy sources such as geothermal, wind, and solar.

Running Out of Energy?

LO17.3 Discuss why running out of oil would not mean running out of energy.

Some observers worry that we may soon run out of the energy needed to power our economy. Their fears are based largely on the possibility that the world may run out of oil sometime in the next century. It is the case, however, that there is no likelihood of running out of energy. If anything, running out of oil would not mean running out of energy—just running out of *cheap* energy.

This is best seen by looking at Table 17.2, which compares oil prices with the prices at which other energy sources become economically viable. For instance, biodiesel, a type of diesel fuel made from decomposed plant wastes, is so expensive to produce that it becomes economically viable (that is, less costly to produce than oil) only if oil costs $80 or more per barrel. On the other hand, ethanol made from corn in the United States costs less to produce and would be an economically viable alternative to oil even if the price of oil were only $60 per barrel.

Multiple Sources of Supply The key point to gather from Table 17.2, however, is that even if we were to run out of oil, alternatives would quickly become available. At a

TABLE 17.2 **Oil Prices at Which Alternative Energy Sources Become Economically Viable**

Oil Price per Barrel at Which Alternative Is Economically Viable	Alternative Fuel
$80	Biodiesel
60	U.S. corn-based ethanol*
50	Shale oil
40	Tar sands; Brazilian sugar-cane-based ethanol; gas to liquids;† coal-to-liquids‡
20	Conventional oil

*Excludes tax credits.

†Gas to liquid is economically viable at $40 if natural gas price is $2.50 or less per million BTUs.

‡Coal to liquid is economically viable at $40 if coal price is $15 per ton or less.

Sources: Cambridge Energy Research Associates, **www.cera.com**; *The Economist*, April 22, 2006, **www.economist.com**.

CONSIDER THIS . . .

Alternative Energy Subsidies and the Fracking Boom

The U.S. government heavily subsidizes alternative energy technologies. Unfortunately, these subsidies have met with many costly failures.

Recent examples include solar-panel maker Solyndra defaulting on over $500 million of federal loan guarantees when it went bankrupt in 2011 and Abound Solar defaulting on $68 million of federal loan guarantees went it went under in 2012. There was an oversupply of solar panels worldwide and prices were too low for many firms to compete—even with heavy subsidies.

The year 2012 also saw the bankruptcies of Ener1 and A123, manufacturers of batteries designed for electric cars. They closed after receiving, respectively, $128 million and over $400 million in government grants and subsidies. They failed because the government had heavily subsidized battery production without considering whether there would be enough demand to keep Ener1, A123, and other battery makers in business. With few electric cars being sold, there was little demand for electric-car batteries.

At the same time, the United States and Canada saw massive increases in oil and natural gas production as the result of hydraulic fracking, which uses injections of superheated steam to release oil and gas from shale formations.

This dramatic change in the supply of conventional energy was confirmed in 2012 by the International Energy Agency, which predicted that fracking would lead to the United States passing Russia as the world's largest natural-gas producer in 2015 and overtaking Saudi Arabia as the world's largest oil producer in 2020.

With conventional energy supplies increasing so massively and alternative energy companies failing so often even when they are subsidized, it is unclear how much political support will remain for subsidizing alternative energy in the future.

Critics contend that the subsidies not only fail to achieve a greater supply of alternative energy but also reflect political corruption that directs government funds to well-connected special interests. Subsidy proponents respond by saying that fossil fuels and fracking are bad for the environment, that alternatives must be developed before the limited supply of fossil fuels runs out, and that conventional energy producers also receive subsidies.

price of $40 per barrel, vast reserves of energy derived from tar sands, the conversion of natural gas and coal to liquid petroleum, and even ethanol derived from cheap Brazilian sugar cane become economically viable alternatives. At $50 per barrel, shale oil becomes a viable alternative. At $60 per barrel, corn-based ethanol becomes viable. And at $80 per barrel, so does biodiesel.

In fact, these prices can be thought of as a giant supply curve for energy, with rising energy prices leading to increased energy production. The result is that even if the supply of oil begins to dry up and oil prices consequently rise, other energy supplies will quickly be brought on line to fill the energy gap created by the decline in the amount of oil available. Also, the alternative prices listed in Table 17.2 are *current* alternative prices. As technologies improve, the costs of producing these alternatives are likely to fall and, as a result, the potential costs of replacing oil if it runs out will be even lower than suggested by the prices in the table. As a result, economists do not worry about running out of oil or, more generally, running out of energy. There is plenty of energy—the only question is price, and the impact of potentially increasing energy prices on the standard of living.

Environmental Impacts Finally, we need to acknowledge that energy sources differ not only in their prices but also with regard to the extent of negative externalities they may generate. Recall from Chapter 4 that negative externalities are costs—such as those associated with air pollution—that are transferred to society during the production process and therefore not reflected in product price. Such externalities need to be accounted for. Some energy sources are relatively "clean," creating little pollution or other externalities. Other sources currently are more problematic. For example, burning coal generates substantial particulate and carbon dioxide emissions that can contribute to health problems and global warming. But a caution is needed here: At sufficiently high electricity prices, burning coal can be both economical and clean. This is true because at sufficiently high electricity prices, the companies that burn coal to generate electricity are able to afford extensive expenditures on pollution reduction. Scrubbers can reduce soot from emissions and new technologies can capture and sequester carbon dioxide in underground storage. At sufficiently high energy prices, clean methods of producing energy are not confined to wind, solar, and other so-called alternative energy sources.

QUICK REVIEW 17.2

- Energy efficiency has consistently improved so that more output can be produced for every unit of energy used by the economy.
- After taking the different fixed costs of different electricity-generating plants into account, utility companies find it efficient to use a variety of energy sources (coal, natural gas, nuclear) to deal with the large daily variations in energy demand with which they must cope.
- Even if we run out of oil, we will not run out of energy because many alternative sources of energy are available. These alternatives are, however, more costly than oil so that if we were to run out of oil, energy costs in the economy would most likely increase.

Natural Resource Economics

LO17.4 Show how the profit motive can encourage resource conservation.

The major focus of natural resource economics is to design policies for extracting or harvesting a natural resource that will maximize the **net benefits** from doing so. The net benefits are simply the total dollar value of all benefits minus the total dollar value of all costs, so that a project's net benefit is equal to the dollar value of the gains or losses to be made. A key feature of such policies is that they take into account the fact that present and future decisions about how fast to extract or harvest a resource typically cannot be made independently. Other things equal, taking more today means having less in the future, and having more in the future is possible only by taking less today.

Renewables vs. Nonrenewables

In applying this general rule, however, large differences between renewable natural resources and nonrenewable natural resources become apparent. **Renewable natural resources** include things like forests and wildlife, which are capable of growing back, or renewing themselves, if they are harvested at moderate rates. This leaves open the possibility of enjoying their benefits in perpetuity. Solar energy, the atmosphere, the oceans, and aquifers are also considered renewable natural resources either because they will continue providing us with their benefits no matter what we do (as is the case with solar energy) or because if we manage them well, we can continue to enjoy their benefits in perpetuity (as is the case with the atmosphere, the oceans, and aquifers). **Nonrenewable natural resources** include things like oil, coal, and metals, which either are in actual fixed supply (like the metals found in the earth's crust) or are renewed so slowly as to be in virtual fixed supply when viewed from a human time perspective (as is the case with fossil fuels like oil and coal, which take millions of years to form out of decaying plants and animals).

Optimal Resource Management

The key to optimally managing both renewable and nonrenewable resources is designing incentive structures that prompt decision makers to consider not only the net benefits to be made by using the resources under their control in the present but also the net benefits to be made by conserving the resources under their control in the present to be able to use more of them in the future. Once these incentive structures are in place, decision makers can weigh the costs and benefits of present use against the costs and benefits of future use to determine the optimal allocation of the resource between present and future uses. The key concept used in weighing these alternatives is present value, which allows decision makers to sensibly compare the net benefits of potential present uses with the net benefits of potential future uses.

Using Present Values to Evaluate Future Possibilities

Natural resource economics studies the optimal use of our limited supplies of resources. Decisions about optimal resource use typically involve choosing how resources will be exploited intertemporally, or over time. For instance, suppose that a poor country has just discovered that it possesses a small oil field. Should the country pump this oil today when it can make a profit of $50 per barrel, or should it wait 5 years to pump the oil given that it believes that in 5 years it will be able to make a profit of $60 per barrel due to lower production costs?

Answering this question requires consideration of the time-value of money, discussed in the previous chapter. We need a way to compare $60 worth of money in 5 years with $50 worth of money today. Economists make this comparison by converting the future quantity of money (in this case $60) into a present-day equivalent measured in present-day money. By making this conversion, the two quantities of money can be compared using the same unit of measurement, present-day dollars.

Understanding Present Values The formula for calculating the present-day equivalent, or **present value,** of any future sum of money (in this case, $60 in 5 years) is described in our macroeconomics chapter on financial economics, but the intuition is simple. Suppose that the current market rate of interest is 5 percent per year. How much money would a person have to save and invest today at 5 percent interest to end up with exactly $60 in 5 years? The correct answer turns out to be $47.01 because if $47.01 is invested at an interest rate of 5 percent per year, it will grow into precisely $60 in 5 years. Stated slightly differently, $47.01 today can be thought of as being equivalent to $60 in 5 years because it is possible to transform $47.01 today into $60 in 5 years by simply investing it at the market rate of interest.

This fact is very important because it allows for a direct comparison of the benefits from the country's two possible courses of action. If it pumps its oil today, it will get $50 per barrel worth of present-day dollars. But if it pumps its oil in 5 years and gets $60 per barrel at that time, it will only get $47.01 per barrel worth of present-day dollars since the present value of $60 in 5 years is precisely $47.01 today. By measuring both possibilities in present-day dollars, the better choice of action becomes obvious: The country should pump its oil today, since $50 worth of present-day money is obviously greater than $47.01 worth of present-day money.

The ability to calculate present values also allows decision makers to use cost-benefit analysis in situations where the costs and benefits happen at different points in time. For instance, suppose that a forestry company is considering spending $1,000 per acre to plant seedlings that it hopes will grow into trees that it will be able to harvest in 100 years. It expects that the wood from the trees will be worth $125,000 per acre in 100 years. Should it undertake this investment? The answer is *no* because at the current market interest rate of 5 percent per year, the present value of $125,000 in 100 years is only $950.56 today, which is less than the $1,000 per acre that the firm would have to invest today to plant the seedlings. When both the benefits and costs of the project are measured in the same units (present-day dollars), it is clear that the project is a money loser and should be avoided.

Allocating Resources over Time More generally, the ability of policymakers to calculate present values and put present-day dollar values on future possibilities is vitally important because it helps to ensure that resources are allocated to their best possible uses over time. By enabling a decision maker to compare the costs and benefits of present use with the costs and benefits of future use, present value calculations help to ensure that a resource will be used at whatever point in time it will be most valuable.

This is especially important when it comes to conservation because there is always a temptation to use up a resource as fast as possible in the present rather than conserving some or all of it for future use. By putting a present-day dollar value on the net benefits to be gained by conservation and future use, present value calculations provide a financial incentive to make sure that resources will be conserved for future use whenever doing so will generate higher net benefits than using them in the present. Indeed, a large part of natural resource economics focuses on nothing more than ensuring that the net benefits that can be gained from conservation and future use are accounted for by the governments, companies, and individuals that are in charge of deciding when and how to use our limited supply of resources. When these future net benefits are properly accounted for, resource use tends to be conservative and sustainable, whereas when they are not properly accounted for, environmental devastation tends to take place, including, as we will discuss in detail below, deforestation and fisheries collapse.

Nonrenewable Resources

Nonrenewable resources like oil, coal, and metals must be mined or pumped from the ground before they can be used. Oil companies and mining companies specialize in the extraction of nonrenewable resources and attempt to make a profit from extracting and then selling the resources that they mine or pump out of the ground. But because extraction is costly and because the price that they will get on the market for their products is uncertain, profits are not guaranteed and such companies must plan their operations carefully if they hope to realize a profit.

User Costs of Current Use We must note, however, that because an oil field or a mineral deposit is typically very large and will take many years to fully extract, an extraction company's goal of "maximizing profits" actually involves attempting to choose an extraction strategy that will maximize a *stream* of profits—potential profits today as well as potential profits in the future. There is, of course, a trade-off. If the company extracts more today, its revenues will be larger today since it will have more product to sell today. On the other hand, more extraction today means that less of the resource will be left in the ground for future extraction and, consequently, future revenues will be smaller since future extraction will necessarily be reduced. Indeed, every bit of resource that is extracted and sold today comes at the cost of not being able to extract it and sell it in the future. Natural resource economists refer to this cost as the **user cost** of extraction because the user of a resource always faces the opportunity cost of reduced future extraction when choosing to extract a resource now rather than in the future.

Present Use versus Future Use The concept of user cost is very helpful in showing how a resource extraction firm that is interested in maximizing its flow of profits over time will choose to behave in terms of how much it will extract in the present as opposed to the future. To give a simple example, consider the case of a coal mining company called Black Rock whose mine will have to shut down in two years, when the company's lease expires. Because the mine will close in two years, the mine's production can be thought of as taking place either during the current year or next year. Black Rock's problem is to figure out how much to mine this year so that it can maximize its stream of profits.

Extraction when Considering Only Current Costs To see how Black Rock's managers might think about the problem, look at Figure 17.7, which shows the situation facing the company during the first year. Begin by noticing *P*, the

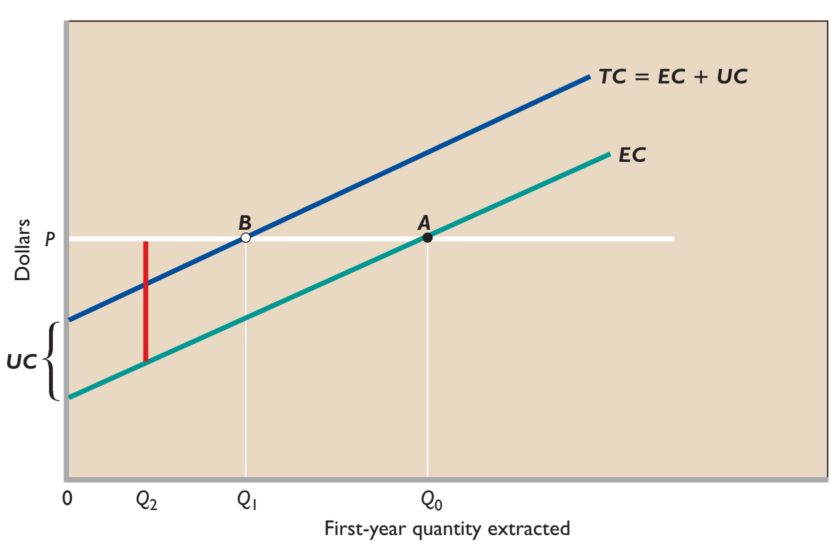


FIGURE 17.7 Choosing the optimal extraction level. A firm that takes account only of current extraction costs, *EC*, will produce Q_0 units of output in the current period—that is, all units for which the market price *P* exceeds extraction costs, *EC*. If it also takes account of user cost, *UC*, and the fact that current output reduces future output and profits, it will produce only Q_1 units of output—that is, only those units for which price exceeds the sum of extraction costs and user cost.

market price at which Black Rock can sell each and every ton of coal that it extracts. The firm's managers will obviously want to take this price into consideration when deciding how much output to produce.

Next, consider the company's production costs, which we will refer to as **extraction costs,** or *EC*, since this is an extraction company. The extraction costs include all costs associated with running the mine, digging out the coal, and preparing the coal for sale. Notice that the *EC* curve that represents extraction costs in Figure 17.7 is upward sloping to reflect the fact that the company's marginal extraction costs increase the more the company extracts because faster extraction involves having to rent or buy more equipment and having to either hire more workers or pay overtime to existing workers. Rapid extraction is costly, and the *EC* curve slopes upward to reflect this fact.

Next, consider how much output the firm's managers will choose to produce if they fail to take user cost into account. If the firm's managers ignore user cost, then they will choose to extract and sell Q_0 tons of coal (shown by where the horizontal *P* line crosses the upward-sloping *EC* line at point *A*). They will do this because for each and every ton of coal that is extracted up to Q_0, the market price at which it can be sold exceeds its extraction cost—making each of those tons of coal profitable to produce.

Extraction When Also Considering User Costs The previous analysis considers only potential first-year profits. None of those tons of coal *has* to be mined this year. Each of them could be left in the ground and mined during the second year. The question that Black Rock's managers have to ask is whether the company's total stream of profits would be increased by leaving some or all of those tons of coal in the ground this year and instead mining and selling them next year.

This question can be answered by taking account of user cost. Specifically, the company's managers can put a dollar amount on how much future profits are reduced by current extraction and then take that dollar amount into account when determining the optimal amount to extract this year. This process is best understood by looking once again at Figure 17.7. There, each ton of coal that is extracted this year is assumed to have a user cost of *UC* dollars per ton that is set equal to the present value of the profits that the firm would earn if the extraction and sale of each ton of coal were delayed until the second year. Taking this user cost into account results in a total cost curve, or *TC*, that is exactly *UC* dollars higher than the extraction cost curve at every extraction level. This parallel upward shift reflects the fact that once the company takes user cost into account, its total costs must be equal to the sum of extraction costs and user cost. That is, $TC = EC + UC$.

If the firm's managers take user cost into account in this fashion, then they will choose to produce less output. In fact, they will choose to extract only Q_1 units of coal (shown by where the horizontal *P* line crosses the upward-sloping *TC* line at point *B*). They will produce exactly this much coal because for each and every ton of coal that is extracted up to Q_1, the market price at which it can be sold exceeds its total cost—including not only the current extraction cost but also the cost of forgone future profits, *UC*.

Why Delay Is More Profitable for Some Units Another way to understand why Black Rock will limit its production to only Q_1 tons of coal is to realize that for every ton of coal up to Q_1, it is more profitable to extract during the current year than during the second year. This is best seen by looking at a particular ton of coal like Q_2. The profit that the firm can get by extracting Q_2 this year is equal to the difference between Q_2's extraction cost and the market price that it can fetch when it is sold. In terms of the figure, this first-year profit is equal to the length of the vertical red line that runs between the point on the *EC* curve above output level Q_2 and the horizontal *P* line.

Notice that the red line is longer than the vertical distance between the *EC* curve and the *TC* curve. This means that the first-year profit is greater than the present value of the second-year profit because the vertical distance between the *EC* curve and the *TC* curve is equal to *UC*, which is by definition the present value of the amount of profit that the company would get if it delayed producing Q_2 until the second year. It is therefore clear that if the firm wants to maximize its profit, it should produce Q_2 during the first year rather than during the second year since the profit to be made by current production exceeds the present value of the profit to be made by second-year production.

This is not true for the tons of coal between output levels Q_1 and Q_0. For these tons of coal, the first-year profit—which is, as before, equal to the vertical distance between the *EC* curve and the horizontal *P* line—is less than *UC*, the present value of the second-year profit that can be obtained by delaying production until the second year. Consequently, the extraction of these units should be delayed until the second year.

Benefits of Efficient Delay The model presented in Figure 17.7 demonstrates that the goal of profit-maximizing extraction firms is not to simply mine coal or pump oil as fast as possible. Instead, they are interested in extracting resources at whatever rate will maximize their streams of profit over time. This incentive structure is very useful to

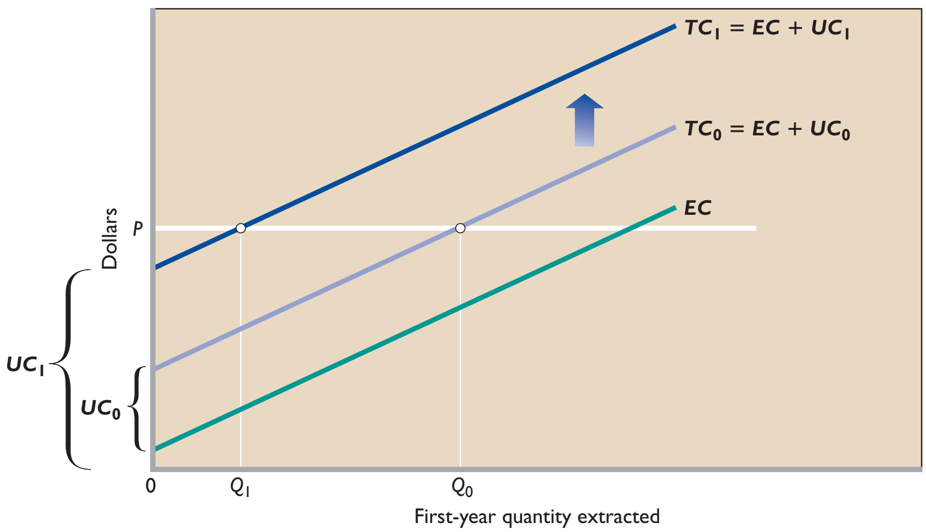


FIGURE 17.8 An increase in expected future profits leads to less current extraction. An increase in future profitability increases user cost from UC_0 to UC_1, thereby raising the total cost curve from TC_0 to TC_1. The firm responds by reducing current production from Q_0 to Q_1 so that it can extract more of the resource in the future and take advantage of the increase in future profitability.

society because it means that our limited supplies of nonrenewable resources will be conserved for future extraction and use if extraction firms expect that demand (and hence profits) in the future will be higher than they are today. This can be seen in Figure 17.8, where user cost has increased in the current period from UC_0 to UC_1 to reflect an increase in expected future profits. This increase in user cost causes Black Rock's total cost curve to shift up from $TC_0 = EC + UC_0$ to $TC_1 = EC + UC_1$. This shift, in turn, reduces the optimal amount of current extraction from Q_0 tons of coal to only Q_1 tons of coal.

This reduction in the amount of coal currently extracted conserves coal for extraction and use in the future when it will be in higher demand. Indeed, Black Rock's profit motive has caused it to reallocate extraction in a way that serves the interests of its customers and their desire to consume more in the future. Since the supply of this nonrenewable resource is limited, more consumption in the future implies less consumption today and Black Rock has accommodated this constraint by reducing extraction this year in order to increase it next year.

More generally speaking, Black Rock's behavior in this case demonstrates that under the right institutional structure, profit-maximizing firms will extract resources efficiently over time, meaning that each unit will tend to be extracted when the net benefits from extraction are the greatest.

Incomplete Property Rights Lead to Excessive Present Use

We just demonstrated that profit-maximizing extraction companies are very happy to decrease current extraction if they can benefit financially from doing so. In particular, they are willing to reduce current extraction if they have the ability to profit from the future extraction and sale of their product. Indeed, this type of a financial situation gives them the incentive to conserve any and all resources that would be more profitably extracted in the future.

This pleasant result breaks down completely if weak or uncertain property rights do not allow extraction companies to profit from conserving resources for future use. For instance, look back at Figure 17.7 and consider how much Black Rock would produce if it were suddenly told that its lease would expire at the end of this year rather than at the end of next year. This would be the equivalent of having a user cost equal to zero because there would be no way for the company to profit in the future by reducing current extraction. As a result, the firm will take into account only current extraction costs, *EC*. The result will be that it will extract and sell Q_0 tons of coal, more than the Q_1 tons that it would extract if it could profit from conservation.

Application: Conflict Diamonds

Resources tend to be extracted much too quickly if there is no way to profit from conservation. That certainly is the case with so-called **conflict diamonds,** which are diamonds mined by combatants in war zones in Africa to provide the hard currency that they need to finance their military activities. Most of these civil wars, however, are very unpredictable, so that control of the mines is tenuous, slipping from one army to another depending on the tide of war.

This fluidity has destroyed any incentive to conserve the resource since the only reason a person would reduce current extraction would be if he or she could benefit from that act of conservation by being able to extract more in the future. But because nobody can be sure of controlling a mine for more than a few months, extraction rates are always extremely high, with the only limit being extraction costs.

This behavior is very wasteful of the resource because once the war finally ends and money is needed to rebuild the country, whichever side wins will find precious few diamonds left to help pay for the reconstruction. Unfortunately, the incentive structures created by the uncertainty of war see to it that extraction takes place at far too rapid a pace, making no allowance for the possibility that future use would be better than present use.

QUICK REVIEW 17.3

- Because nonrenewable resources are finite, it is very important to allocate their limited supply efficiently between present and future uses.
- If resource extraction companies can benefit from both present and future extraction, they will limit current extraction to only those units that are more profitable to extract in the present rather than in the future. This conserves resources for future use.
- If resource users have no way of benefiting from the conservation of a resource, they will use too much of it in the present and not save enough of it for future use—even if future use would be more beneficial than present use.

Renewable Resources

LO17.5 Relate how to use property rights to prevent deforestation and species extinction.

We just saw that under the right circumstances, extraction companies have a strong profit incentive to moderate their current extraction rates and conserve nonrenewable resources for future use. A similar incentive can also hold true for companies and individuals dealing with renewable resources like forests and wildlife. If property rights are structured properly, then decision makers will have an incentive to preserve resources and manage them on a sustainable basis, meaning that they will harvest the resources slowly enough that the resources can always replenish themselves.

On the other hand, if proper incentives are not in place, then high and nonsustainable harvest rates can quickly wipe out a renewable resource. Indeed, ecologists and natural resource economists can cite numerous examples of fish and animal populations collapsing because of overfishing and overhunting, as well as rainforests being wiped out because of overlogging. This section discusses the economics of renewable resources as well as policies that promote the sustainable use of renewable resources. To keep things concrete, we provide a quick example of overhunting and then turn our main attention to forests and fisheries.

Elephant Preservation

If a renewable wildlife resource is used too fast, it can become extinct. This was the situation facing elephants in Africa during the 1970s and 1980s when elephant populations in most parts of Africa declined drastically due to the illegal poaching of elephants for their ivory tusks. It was the case, however, that elephant populations in a few countries expanded considerably. The difference resulted from the fact that in certain countries like Botswana and Zimbabwe, property rights over elephants were given to local villagers, thereby giving them a strong financial incentive to preserve their local elephant populations. In particular, local villagers were allowed to keep the money that could be earned by taking foreign tourists on safari to see the elephants in their area as well as the money that could be made by selling hunting rights to foreign sports hunters. This gave them a strong incentive to prevent poaching, and villagers quickly organized very effective patrols to protect and conserve their valuable resource.

By contrast, elephants belonged to the state in other countries, meaning that locals had no personal stake in the long-term survival of their local elephant populations since any elephant tourism money flowed to the state and other outsiders. This created the perverse incentive that the only way for a local to benefit financially from an elephant was by killing it to get its ivory. Indeed, most of the poaching in these countries was done by locals who had been given no way to benefit from the long-term survival of their local elephant populations. As with nonrenewable resources, the inability to benefit from conservation and future use causes people to increase their present use of renewable resources.

Forest Management

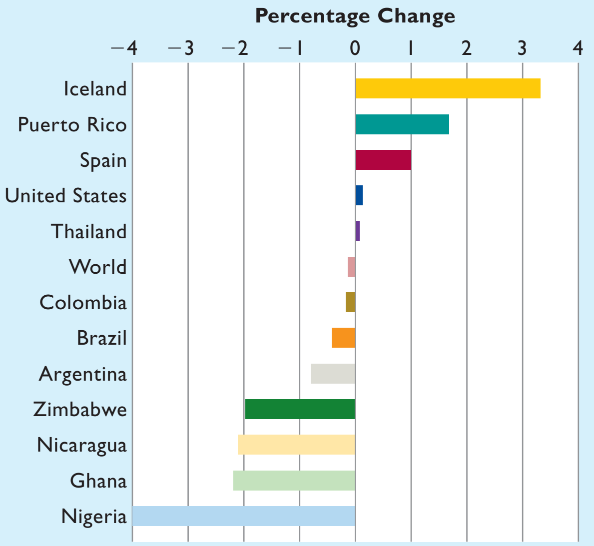
Forests provide many benefits including wildlife habitat, erosion prevention, oxygen production, recreation, and, of course, wood. In 2010, just under 10 billion acres, or about 30 percent of the world's land area, was forested and about 746 million acres, or about 33 percent of the United States' land area, was forested. The amount of land covered by forests is, however, growing in some places but declining in others. This fact is apparent in Global Perspective 17.1, which gives the total percentage change over the years 2005 to 2010 in the amount of forest-covered land in 12 selected countries as well as in the entire world.

The Importance of Property Rights Economists believe that the large variation in growth rates seen in

GLOBAL PERSPECTIVE 17.1

Percentage Change in the Amount of Land Covered by Forests, 2005–2010

The percentage change in the amount of land covered by forests varies greatly by nation.



Source: Global Forests Resource Assessment 2010, Food and Agriculture Organization of the United Nations, **www.fao.org**. Used with permission.

Global Perspective 17.1 is largely the result of differences in property rights. In certain areas, including the United States and western Europe, forests are either private property or strictly regulated government property. In either case, individuals or institutions have an incentive to harvest their forests on a sustainable basis because they can benefit not just from cutting down the trees currently alive but also from keeping their forests going to reap the benefits that they will give off in the future if they are managed on a sustainable basis.

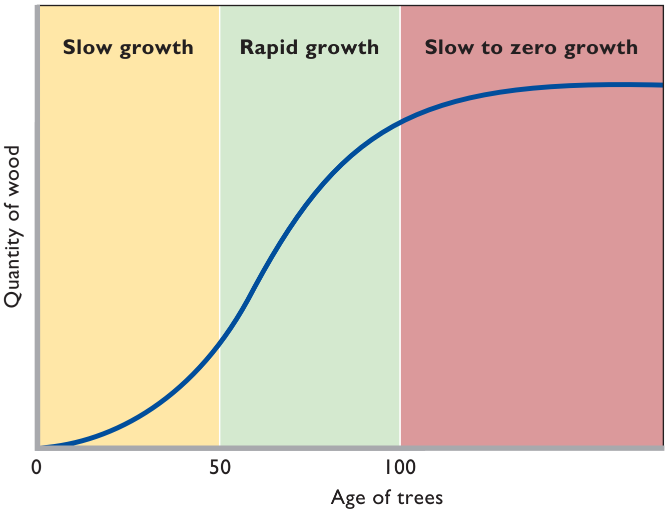
By contrast, deforestation is proceeding rapidly in countries where property rights over forests are poorly enforced or nonexistent. To see why this is true, consider the situation facing competing loggers if nobody owns the property rights to a given forest. In such a situation, whoever chops down the forest first will be able to reap economic benefits because, while nobody can have ownership or control over a living tree, anybody can establish a property right to it by chopping it down and bringing it to market. In such a situation, everybody has an incentive to chop down as many trees as fast as they can to secure them before anyone else can. Sadly, nobody has an incentive to preserve trees for future use because—without enforceable property rights—person A has no way to prevent person B from chopping down the trees that person A would like to preserve.

To reduce and hopefully eliminate nonsustainable logging, governments and international agencies have been taking increasingly strong steps to define and enforce property rights over forests. One major result is that in areas such as the United States and Europe where strong property rights over forests have been established, virtually all wood production is generated by commercially run forestry companies. These companies buy up large tracts of land on which they plant and harvest trees. Whenever a harvest takes place and the trees in a given area are chopped down, seedlings are planted to replace the felled trees, thereby replenishing the stock of trees. These companies are deeply concerned about the long-term sustainability of their operations and many often plant trees in the expectation that more than a century may pass before they are harvested.

Optimal Forest Harvesting In cases where the property rights to a forest are clear and enforceable (as they are in the United States), forest owners have a strong incentive to manage their forests on a sustainable basis because they can reap the long-term benefits that derive from current acts of conservation. A key part of their long-term planning is deciding how often to harvest and then replant their trees.

This is an interesting problem because a commercial forestry company that grows trees for lumber or paper production must take into consideration the fact that trees grow at different rates over the course of their lifetimes. Indeed, Figure 17.9 shows that if the company plants an

FIGURE 17.9 A forest's growth rate depends on its age. Because trees do not reach their most rapid growth rates until middle age, forestry companies have an incentive not to harvest them too early. But because growth then tapers off as the trees reach their maximum adult sizes, there is an incentive to cut them down before they are fully mature to enable the replanting of the forest with faster-growing young trees.



acre of land with seedlings and lets those seedlings grow into mature trees, the amount of wood contained in the trees at first grows rather slowly as the seedlings slowly grow into saplings, then grows quite quickly as the saplings mature into adult trees, and then tapers off as the adult trees reach their maximum size.

This growth pattern leads forestry companies to think very carefully about when to harvest their trees. If they harvest and replant the acre of land when the trees are only 50 years old, they will miss out on the most rapid years of growth. On the other hand, there is not much point in letting the trees get much more than 100 years old before harvesting and replanting since at that age very little growth is left in them. The result is that the forestry company will choose to harvest the trees and replant the land when the trees reach an age of somewhere between 50 and 100 years old. The precise age will be chosen to maximize firm profits and will be affected not only by the growth rate of trees but also by other factors including the cost of harvesting the trees and, of course, the market price of wood and how it is expected to vary over time.

The key point to keep in mind, however, is that forestry companies that have secure property rights over their trees do not harvest them as soon as possible. Instead, they shepherd their resource and harvest their trees only when replacing older, slow-growing trees with younger, fast-growing trees finally becomes more profitable. And, of course, it must also be emphasized that forestry companies *replant*. They do this because they know that they can benefit from the seedlings' eventual harvest, even if that is 50 or 100 years in the future. In countries where property rights are not secure, nobody replants after cutting down a forest because there is no way to prevent someone else from cutting down the new trees and stealing the harvest.

Optimal Fisheries Management

A **fishery** is a stock of fish or other marine animals that can be thought of as a logically distinct group. A fishery is typically identified by location and species—for example, Newfoundland cod, Pacific tuna, or Alaskan crab. Table 17.3 lists the top 10 U.S. fisheries in terms of how much their respective catches were worth in 2010.

The key difficulty with fishery management is that the only way to establish property rights over a fish swimming in the open ocean is to catch it and kill it. As long as the fish is alive and swimming in the open ocean, it belongs to nobody. But as soon as it is caught, it belongs to the person who catches it. This property rights system means that the only way to benefit economically from a fish is to catch it and thereby turn it into a private good.

TABLE 17.3 **Top 10 U.S. Fisheries in Dollar Terms, 2010**

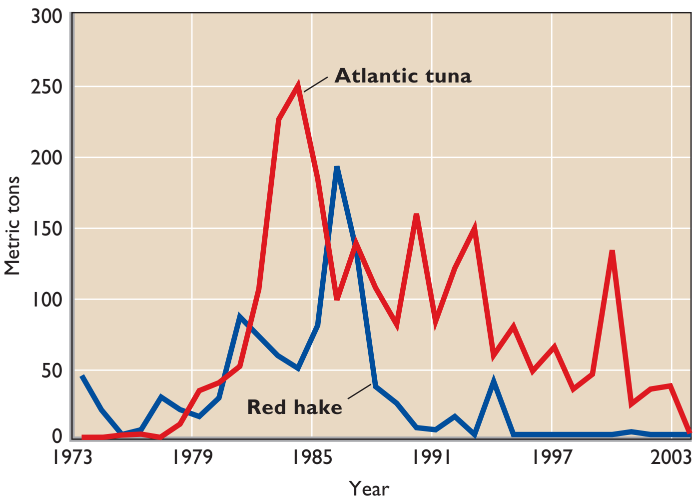
Fishery	Market Value of Catch
Sea scallop	$455,693,743
Lobster	399,476,190
Walleye pollock	282,399,223
Sockeye salmon	278,646,491
Pacific halibut	206,958,364
White shrimp	191,608,324
Blue crab	189,784,233
Pacific cod	146,940,754
Brown shrimp	144,592,574
Dungeness crab	140,443,133

Source: National Ocean Economics Program, **noep.mbari.org**. Information provided by Judith Kildow and the NOEP team at the Monterey Bay Aquarium Research Institute.

This creates an incentive for fishers to be very aggressive and try to outfish each other, since the only way for them to benefit from a particular fish is to catch it before someone else does. The calamitous result of this perverse incentive has been tremendous overfishing, which has caused many fisheries to collapse and which threatens many others with collapse as well.

Two examples of fishery collapse are presented in Figure 17.10, which shows the number of metric tons per year of Maine red hake and Atlantic tuna that were caught between 1973 and 2004 by U.S. fishers. A **fishery collapse** happens when a fishery's population is sent into a rapid decline because fish are being harvested faster than they can reproduce. The speed of the decline depends on how much faster harvesting is than reproduction. In the case of Maine red hake, the decline was very abrupt, with the annual catch falling from 190.3 million metric tons in 1986 down to only 4.1 million tons 5 years later. After making a minor resurgence in 1994, the fishery then totally collapsed, so that the catch was less than 1 ton per year for most of the following decade despite the best efforts of fishers to catch more. The collapse of the Atlantic tuna fishery was more gradual, presumably because the ratio of harvest to reproduction was not as extreme as it was for Maine red hake. But even when harvesting exceeds reproduction by only a small amount in a given year, the population declines. And if that pattern holds for many years, the population will be forced into collapse. This was the case for Atlantic tuna. Its annual catch collapsed more gradually, from a peak of 248.9 million metric tons in 1984 down to only 4.1 million metric tons in 2004.

Overfishing and fishery collapse are now extremely common, so much so that worldwide stocks of large predatory fish like tuna, halibut, swordfish, and cod are believed to be



Source: National Marine Fisheries Service, National Oceanic and Atmospheric Administration, **www.nmfs.noaa.gov**.

FIGURE 17.10 The collapse of two fisheries, 1973–2004. This figure shows how many metric tons of Atlantic tuna and Maine red hake were caught by U.S. fishing boats each year from 1973 to 2004. Overfishing has caused the population of both species to collapse, Maine red hake very abruptly and Atlantic tuna more slowly.

90 percent smaller than they were just 50 years ago. In addition, Table 17.4 shows that just 13 percent of world fisheries in 2009 were estimated to be underexploited, whereas nearly 60 percent were categorized as fully exploited, and 30 percent were categorized as overexploited, meaning they were either depleted or (hopefully) recovering from depletion.

Policies to Limit Catch Sizes

Governments have tried several different polices to limit the number of fish that are caught each year to prevent fisheries from collapsing. They also hope to lower annual catch sizes down to sustainable levels, where the size of the catch does not exceed the fishery's ability to regenerate. Unfortunately, many of these policies not only fail to reduce catch sizes but also create perverse incentives that raise fishing costs because they do not stop the fishing free-for-all in which each fisher tries to catch as many fish as possible as fast as possible before anyone else can get to them.

Shortening the Length of Fishing Seasons For example, some policies attempt to reduce catch sizes by limiting the number of days per year that a certain species can be caught. For instance, the duration of the legal crabbing season in Alaska was once cut down from several months to just 4 days. Unfortunately, this policy failed to reduce catch sizes because crabbers compensated for the short legal crabbing season by buying massive boats that could harvest in 4 days the same amount of crab that they had previously needed months to gather.

Fishers bought the new, massive boats because while the new policy limited the number of days over which crabbers were allowed to compete, it did not lessen their incentive to try to catch as many crabs as possible before anyone else could get to them. Indeed, the massive new boats were a sort of arms race, with each fisher trying to buy a bigger, faster, more powerful boat than his competitors to be able to capture more of the available crabs during the limited 4-day season. The result, however, was a stalemate because if everybody is buying bigger, faster, more powerful boats, then nobody gains an advantage. Consequently, the policy actually made the situation worse. Not only did it fail to reduce catch size; it also drove up fishing costs. This was an especially pernicious result because the policy had been designed to help fishers by preserving the resource upon which their livelihoods depended.

TABLE 17.4 Status of World's Fisheries in 2009

Status	Percentage
Nonfully exploited	13%
Fully exploited	57
Overexploited	30

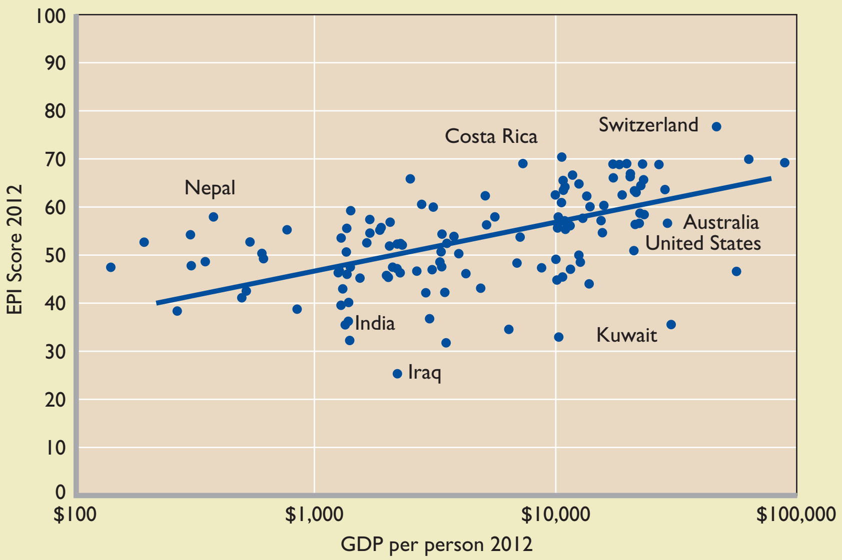
Source: State of the World's Fisheries and Aquaculture, 2012, Food and Agriculture Organization of the United Nations, **www.fao.org**. Used with permission.

Limiting the Number of Boats Another failed policy attempted to limit catch size by limiting the number of fishing boats allowed to fish in a specific area. This policy failed because fishers compensated for the limit on the number of boats by operating bigger boats. That is, many small boats that could each catch only a few tons of fish were replaced by a few large boats that could each catch many tons of fish. Once again, catch sizes did not fall.

LAST WORD

Is Economic Growth Bad for the Environment?

Measures of Environmental Quality are Higher in Richer Countries.



Many people are deeply concerned that environmental degradation is an inevitable consequence of economic growth. Their concern is lent credence by sensational media events like oil and chemical spills and by the indisputable fact that modern chemistry and industry have created and released into the environment many toxic chemicals and substances that human beings did not even know how to make a couple of centuries ago.

Economists, however, tend to be rather positive about economic growth and its consequences for the environment. They feel this way because significant evidence indicates that richer societies spend much more money on keeping their respective environments healthy than do poorer societies. Viewed from this perspective, economic growth and rising living standards are good for the environment because as societies get richer, they tend to spend more on things like reducing emissions from smokestacks, preventing the dumping of toxic chemicals, and insisting that sewage be purified before its water is returned to the environment. They also tend to institute better protections for sensitive ecosystems and engage in greater amounts of habitat preservation for endangered species.

But are these increasing expenditures on environmentally beneficial goods and services enough to overcome the massive increases in environmental harm that seem likely to accompany the enormous amounts of production and consumption in which rich societies engage? The empirical record suggests that the

Limiting the Total Catch A policy that does work to reduce catch size goes by the acronym **TAC,** which stands for **total allowable catch.** Under this system, biologists determine the TAC for a given fishery, for instance, 100,000 tons per year. Fishers can then fish until a total of 100,000 tons have been brought to shore. At that point, fishing is halted for the year.

This policy has the benefit of actually limiting the size of the catch to sustainable levels. But it still encourages an arms race between the fishers because each fisher wants to try to catch as many fish as possible before the TAC limit is reached. The result is that even under a TAC, fishing costs rise because fishers buy bigger, faster boats as each one tries to fulfill as much of the overall TAC catch limit as possible.

Assigning Individual Transferable Quotas The catch-limiting system that economists prefer not only limits the total catch size but also eliminates the arms race between fishers that drives up costs. The system is based on the issuance of **individual transferable quotas,** or **ITQs,** which are individual catch size limits that specify that the holder of an ITQ has the right to harvest a given quantity of a particular species during a given time period, for instance, 1,000 tons of Alaskan king crab during the year 2017.

The individual catch sizes of all the ITQs that are issued for a given fishery during a specific year add up to the fishery's overall TAC for the year so that they put a sustainable limit on the overall catch size. This preserves the fishery from overexploitation. But the fact that the ITQ quotas are *individual* also eliminates the need for an arms race. Because each fisherman knows that he can take as long as he wants to catch his individual quota, he does not need a superexpensive, technologically sophisticated boat that is capable of hauling in massive amounts of fish in only a few days in order to beat his competitors to the punch. Instead, he can use smaller, less expensive, and answer is yes. The best evidence for this is given by the accompanying figure, in which each of 125 countries is represented by a point that indicates both its GDP per capita (measured on the horizontal axis using a logarithmic scale) and its year 2012 score on the Environmental Performance Index, or EPI.

This index, produced by researchers at Yale University, compares countries based on how well they are doing in terms of 25 environmental indicators, including atmospheric carbon emissions, measures of air and water quality, the degree of wilderness protection, energy efficiency, and measures of whether a country's fisheries and forests are being overexploited. Out of a maximum possible EPI score of 100, Switzerland and Latvia received the highest scores of, respectively, 76.69 and 70.37. The United States was ranked 49th with a score of 56.59 while the lowest-ranked country, Iraq, received a score of 25.32.

When EPI scores are combined with measures of GDP per person in the figure, an extremely strong pattern emerges: Richer countries have higher EPI scores. In fact, the relationship between the two variables is so strong that 70 percent of the differences between countries in terms of EPI scores are explained by their differences in GDP per person. In addition, the logarithmic scale used on the horizontal axis allows us to look at the best-fit line drawn through the data and conclude that a 10-fold increase in GDP per capita (from, for instance, \$1,000 to \$10,000) is associated with a 10-point increase in EPI. The figure is therefore clear confirmation that economic growth can not only go together with a healthy environment, but that economic growth actually promotes a healthy environment by making people rich enough to pay for pollution-reduction technologies that people living in poorer countries cannot afford.

Looking to the future, many economists are hopeful that economic growth and rising living standards will pay for the invention and implementation of new technologies that could make for an even cleaner environment. If the current pattern continues to hold, increased standards of living will lead to better environmental outcomes.

Note: The horizontal axis is measured using a logarithmic scale, so that each successive horizontal unit represents a 10-fold increase in GDP per person. This is useful because it happens to be the case that the relationship between EPI and GDP per person is such that a 10-fold increase in GDP per person is associated with a 20-point increase in EPI. Graphing the data using a logarithmic scale makes this relationship obvious.

Sources: The EPI data as well as the purchasing-power-parity-adjusted per-person GDP data are from the Yale Center for Environmental Law and Policy (YCELP) and Center for International Earth Science Information Network (CIESIN), **epi.yale.edu**.

simpler boats since he knows that he can fish slowly—perhaps year round if it suits him.

Efficiency Gains This move toward smaller boats and more leisurely fishing greatly reduces fishing costs. But ITQs offer one more cost-saving benefit. They encourage all of the fishing to be done by the lowest-cost, most-efficient fishing vessels. This is true because ITQs are *tradable* fishing quotas, meaning that they can be sold and thereby traded to other fishers. As we will explain, market pressures will cause them to be sold to the fishers who can catch fish most efficiently, at the lowest possible cost.

To see how this works, imagine a situation in which the market price of tuna is \$10 per ton but in which a fisherman named Sven can barely make a profit because his old, slow boat is so expensive that it costs him \$9 per ton to catch tuna. At that cost, if he does his own fishing and uses his ITQ quota of 1,000 tons himself, he will make a profit of only \$1,000 (= \$1 per ton × 1,000 tons). At the same time, one of his neighbors, Tammy, has just bought a new, superefficient ship that can harvest fish at the very low cost of \$6 per ton. This difference in costs means that Sven and Tammy will both find it advantageous to negotiate the sale of Sven's ITQ to Tammy. Sven, for his part, would be happy to accept any price higher than \$1,000 since \$1,000 is the most that he can make if he does his own fishing. Suppose that they agree on a price of \$2 per ton, or \$2,000 total. In such a case, both are better off. Sven is happy because he gets \$2,000 rather than the \$1,000 that he would have earned if he had done his own fishing. And Tammy is happy because she is about to make a tidy profit. The 1,000 tons of tuna that she can catch with Sven's ITQ will bring \$10,000 in revenues when they are sold at \$10 per ton, while her costs of bringing in that catch will be only \$8,000 since it costs \$6,000 in fishing costs at \$6 per ton to bring in the catch plus the \$2,000 that she pays to Sven for the right to use his 1,000-ton ITQ.

CONSIDER THIS . . .

The Tragedy of the Commons

In an article titled "The Tragedy of the Commons," ecologist Garret Hardin explained the crucial role that individual property rights play in resource preservation.

Hardin discussed the public plots of grazing land that were set aside in many villages in medieval Europe. These plots were called commons, after the fact that they were held in common and could be used by anyone. They were a form of welfare designed to help poor people graze and feed animals even if they couldn't afford any land of their own.

Hardin pointed out that this welfare system often failed due to a lack of individual property rights. In particular, the commons were overrun, overgrazed, and turned into barren patches of dirt because they were "first come, first served." The fact that anybody could use the commons meant that nobody had an individual incentive to try to preserve an existing patch of grass. That was because any grass that one person chose to preserve would just end up being eaten by somebody else's animals. So if a person saw any uneaten grass, his best strategy was to let his animals devour it before somebody else's animals did.

As soon as Hardin published his article, people realized that similar **tragedy of the commons** situations were prone to occur wherever individual property rights were lacking. Consider overfishing and deforestation. They both occur because a lack of individual property rights means that each user is incentivized to use as much as possible, as quickly as possible, before anyone else can get to the resource.

Social Benefits Notice, though, that society also benefits. If Sven had used his ITQ himself, he would have run up fishing costs of $9,000 harvesting the 1,000 tons of tuna that his quota allows. But because the permit was sold to Tammy, only $6,000 in fishing costs are actually incurred. The tradable nature of ITQs promotes overall economic efficiency by creating an incentive structure that tends to move production toward the producers who have the lowest production costs.

It remains to be seen, however, if ITQs and other catch-reduction policies will be enough to save the world's fisheries. Since current international law allows countries to enforce ITQs and other conservation measures only within 200 miles of their shores, most of the world's oceans are a fishing free-for-all. Unless this changes and incentive structures are put in place to limit catch sizes in international waters, economic theory suggests that the fisheries there will continue to decline as fishers compete to catch as many fish as possible as fast as possible before anyone else can get to them.

QUICK REVIEW 17.4

- When property rights are absent, renewable resources tend to be depleted quickly because users have no way of benefiting from conservation.
- Governments that establish and enforce property rights over renewable resources encourage conservation by allowing users to benefit financially from future harvesting as well as present harvesting.
- Total allowable catch limits (TAQs) combined with individual transferrable quotas (ITQs) promote the preservation and efficient harvesting of fisheries. The TAQs preserve fisheries by capping total harvest sizes. The ITQs promote efficiency by providing financial incentives that encourage all the fishing to be done by the most efficient fishers.

SUMMARY

LO17.1 Explain why falling birthrates mean that we are not likely to run out of natural resources.

Per capita living standards in the United States are at least 12 times higher than they were in 1800. This increase in living standards has entailed using much larger amounts of resources to produce the much larger amounts of goods and services that are currently consumed. The increase in resource use can be attributed to two factors. First, there has been a large increase in resource use per person. Second, there are now many more people alive and consuming resources than at any previous time.

The large increase in total resource use has led to a spirited debate about whether our high and rising living standards are sustainable. In particular, will our demand for resources soon outstrip the supply of resources? A proper answer to this question involves examining the demand for resources as well as the supply of resources.

A good way to examine the demand for resources is to think of total resource demand as being the product of the amount of resources used per person times the number of people alive. Thomas Malthus famously predicted that higher living standards would tend to lead to higher birthrates. The opposite, however, has held true. Higher living standards have led to lower birthrates and the majority of the world's population now lives in countries where the total fertility rate is less than the replacement rate of 2.1 births per woman per lifetime necessary to keep a country's population stable over time.

The result is that world population growth is not only slowing but is actually turning negative in many countries. What is more, the effect of low birthrates is so strong that many demographers believe that the world's population will reach a maximum of fewer than 9 billion people in the next 50 years before beginning to decline quite rapidly. That implies substantially reduced resource demand.

The evidence from the United States and other rich countries is that resource use per person has either fallen or leveled off during the past several decades. For instance, per capita water use in the United States fell 28 percent between 1975 and 2000. Per capita energy use has declined since the late 1980s. And because the per capita generation of trash has been stable since 1990, we can infer that the per capita use of solid objects like metals, paper, and plastics has been stable since that time as well.

Combined with the expected decline in population levels, the fact that per capita resource use has either fallen or leveled off implies that the total demand for resources is likely to reach a peak in the next 50 years before falling over time as populations decline.

Natural resource economists predict that resource supplies are likely to grow faster than resource demands in the future. This confidence is based on the fact that since 1850 the real (inflation-adjusted) prices of resources have fallen by about 50 percent. Because this decline in prices happened at the same time that total resource use was increasing dramatically, it seems likely that resource supplies will continue to grow faster than resource demands since, going forward, resource use should grow less quickly than it has in the past because population growth has slowed (and is expected to turn negative) and because per capita resource use in recent decades has leveled off or turned negative.

LO17.2 Describe why using a mix of energy sources is efficient, even if some of them are quite costly.

Living standards can continue to rise without consuming more energy thanks to more efficient productive technologies, which can produce more output using the same amount of energy input. Indeed, real GDP per person in the United States increased by nearly 40 percent between 1988 and 2011 despite the fact that annual per capita energy consumption fell 8 percent during those years.

Differences in fixed costs mean that a wide variety of energy sources are used in the economy despite the fact that some of these energy sources are much more costly than others. For instance, coal-fired electric generating plants use low-cost coal, but are extremely expensive to build so that they are used only in situations where very large generating capacities are required. By contrast, when smaller amounts of electricity are required, it often makes more sense to employ other generating technologies such as natural gas even though they use more expensive fuel.

LO17.3 Discuss why running out of oil would not mean running out of energy.

We are not running out of energy. Even if we run out of oil, there are plenty of other energy sources including biodiesel, ethanol made from corn or sugar cane, and oil made from organic waste products. The only question is cost.

LO17.4 Show how the profit motive can encourage resource conservation.

Renewable natural resources like forests and fisheries as well as nonrenewable natural resources like oil and coal tend to be overused in the present unless there are institutions created that provide resource users with a way to benefit from conservation. Governments can ensure this benefit by strictly defining and enforcing property rights so that users know that if they conserve a resource today, they will be able to use it or sell it in the future.

LO17.5 Relate how to use property rights to prevent deforestation and species extinction.

Encouraging conservation is especially difficult in the open ocean where it is impossible to either define or enforce property rights over fish because, by international law, nobody owns the open ocean and so anyone can fish there as much as he or she wants. This lack of property rights leads to severe overfishing and an eventual collapse of the fishery.

Closer to shore, however, governments can define property rights within their sovereign waters and impose limits on fishing. The best system involves combining total allowable catch (TAC) limits for a given fishery with individual transferable quotas (ITQs) for individual fishers.

TERMS AND CONCEPTS

replacement rate
total fertility rate
demographers
demographic transition
British thermal unit (BTU)
net benefits
renewable natural resources
nonrenewable natural resources
present value
user cost
extraction cost
conflict diamonds
fishery
fishery collapse
total allowable catch (TAC)
individual transferable quotas (ITQs)
tragedy of the commons

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Describe Thomas Malthus's theory of human reproduction. Does it make sense for some species—say, bacteria or rabbits? What do you think makes humans different? **LO17.1**
2. Demographers have been surprised that total fertility rates have fallen below 2.0, especially because most people in most countries tell pollsters that they would like to have at least two children. Can you think of any possible economic factors that may be causing women in so many countries to average fewer than two children per lifetime? What about other social or political changes? **LO17.1**
3. Resource consumption per person in the United States is either flat or falling, depending on the resource. Yet living standards are rising because of technological improvements that allow more output to be produced for every unit of input used in production. What does this say about the likelihood of our running out of resources? Could we possibly maintain or improve our living standards even if the population were expected to rise in the future rather than fall? **LO17.1**
4. A community has a nighttime energy demand of 50 megawatts but a peak daytime demand of 75 megawatts. It has the chance to build a 90-megawatt coal-fired plant that could easily supply all of its energy needs even at peak daytime demand. Should it necessarily proceed? Could there be lower-cost options? Explain. **LO17.2**
5. Suppose that you hear two people arguing about energy. One says that we are running out of energy. The other counters that we are running out of cheap energy. Explain which person is correct and why. **LO17.3**
6. Recall the model of nonrenewable resource extraction presented in Figure 17.7. Suppose that a technological breakthrough means that extraction costs will fall in the future (but not in the present). What will this do to future profits and, therefore, to current user cost? Will current extraction increase or decrease? Compare this to a situation where future extraction costs remain unchanged but current extraction costs fall. In this situation, does current extraction increase or decrease? Does the firm's behavior make sense in both situations? That is, does its response to the changes in production costs in each case maximize the firm's stream of profits over time? **LO17.4**
7. If the current market price rises, does current extraction increase or decrease? What if the future market price rises? Do these changes in current extraction help to ensure that the resource is extracted and used when it is most valuable? **LO17.4**
8. **ADVANCED ANALYSIS** Suppose that a government wants to reduce its economy's dependence on coal and decides as a result to tax coal mining companies $1 per ton for every ton of coal that they mine. Assuming that coal mining companies treat this tax as an increase in extraction costs this year, what effect will the tax have on current extraction in the model used in Figure 17.7? Now, think one step ahead. Suppose that the tax will be in place forever, so that it will also affect extraction costs in the future. Will the tax increase or decrease user cost? Does this effect increase or decrease the change in current extraction caused by the shift of the EC curve? Given your finding, should environmental taxes be temporary? **LO17.4**
9. **ADVANCED ANALYSIS** User cost is equal to the present value of future profits in the model presented in Figure 17.7. Will the optimal quantity to mine in the present year increase or decrease if the market rate of interest rises? Does your result make any intuitive sense? (Hint: If interest rates are up, would you want to have more or less money right now to invest at the market rate of interest?) **LO17.4**
10. Various cultures have come up with their own methods to limit catch size and prevent fishery collapse. In old Hawaii, certain fishing grounds near shore could be used only by certain individuals. And among lobstermen in Maine, strict territorial rights are handed out so that only certain people can harvest lobsters in certain waters. Discuss specifically how these systems provide incentives for conservation. Then think about the enforcement of these property rights. Do you think similar systems could be successfully enforced for deep-sea fishing, far off shore? **LO17.5**
11. Aquaculture is the growing of fish, shrimp, and other seafood in enclosed cages or ponds. The cages and ponds not only keep the seafood from swimming away but also provide aquaculturists with strong property rights over their animals. Does this provide a good incentive for low-cost production as compared with fishing in the open seas where there are few if any property rights? **LO17.5**
12. **LAST WORD** The figure in the Last Word section shows that a 10-fold increase in a country's GDP per person is associated with about a 20-point increase in EPI. On the other hand, GDP per person was $48,112 in the United States in 2011 but $36,254 in New Zealand; yet New Zealand had an EPI score of 66.05, while the United States had an EPI score of only 56.59. So does getting rich guarantee doing well environmentally? Discuss.

REVIEW QUESTIONS

1. The long-run downward trend in commodity prices is consistent with the idea that: **LO17.1**
   a. We are quickly running out of resources.
   b. Resource demands have been increasing faster than resource supplies.
   c. Birthrates will soon increase due to the falling cost of living.
   d. Resource supplies have increased faster than resource demands.

2. It would cost the town of Irondale $50 million to build a gas-powered generator that could produce a maximum of 5 megawatts of electricity at 15 cents per hour. Another alternative would be for Irondale to build a $100 million coal-fired generator that could produce a maximum of 15 megawatts of electricity at 5 cents per hour. Irondale should: **LO17.2**
   a. Build the coal-fired generator because its hourly operating costs are so much lower.
   b. Build the gas-powered generator since it is less expensive to build.
   c. Build the coal-fired generator because, while it would cost twice as much to build, it would produce three times as much electricity.
   d. Obtain more information before deciding what to do.
3. After mining 9,273 tons of coal, Blue Sky Mining's managers note that the marginal cost of mining the next ton of coal would be $40 per ton. They also calculate that the user cost of mining that next ton of coal would be $35. If the market price of coal is $72, should Blue Sky mine an additional ton of coal? **LO17.4**
   a. Yes.
   b. No.
   c. More information is needed.
4. Good methods for helping to protect natural resources include: **LO17.5**
   a. Establishing property rights and giving them to local users.
   b. Encouraging first-come, first-served property rights.
   c. Teaching people to consider user cost.
   d. Having the government set up and enforce ITQs.
5. Ingvar and Olaf are the only two fishermen in their area. Each has been assigned an ITQ that allows him to catch 20 tons of salmon. Ingvar's MC of catching salmon is $6 per ton while Olaf's MC of catching salmon is $7 per ton. If the price of salmon is $10 per ton, then to maximize efficiency, the two guys should trade ITQs until Ingvar is in charge of catching ________ tons while Olaf catches ________ tons. **LO17.5**
   a. 20; 20.
   b. 30; 10.
   c. 40; 0.
   d. 0; 40.

PROBLEMS

1. Suppose that the current (first) generation consists of 1 million people, half of whom are women. If the total fertility rate is 1.3 and the only way people die is of old age, how big will the fourth generation (the great-grandchildren) be? How much smaller (in percentage terms) is each generation than the previous generation? How much smaller (in percentage terms) is the fourth generation than the first generation? Are you surprised by how quickly the population declines? **LO17.1**
2. A coal-fired power plant can produce electricity at a variable cost of 4 cents per kilowatt hour when running at its full capacity of 30 megawatts per hour, 16 cents per kilowatt hour when running at 20 megawatts per hour, and 24 cents per kilowatt hour when running at 10 megawatts per hour. A gas-fired power plant can produce electricity at a variable cost of 12 cents per kilowatt-hour at any capacity from 1 megawatt per hour to its full capacity of 5 megawatts per hour. The cost of constructing a coal-fired plant is $50 million, but it costs only $10 million to build a gas-fired plant. **LO17.2**
   a. Consider a city that has a peak afternoon demand of 80 megawatts of electricity. If it wants all plants to operate at full capacity, what combination of coal-fired plants and gas-fired plants would minimize construction costs?
   b. How much will the city spend on building that combination of plants?
   c. What will the average cost per kilowatt-hour be if you average over all 80 megawatts that are produced by that combination of plants? (Hint: A kilowatt is one thousand watts, while a megawatt is one million watts).
   d. What would the average cost per kilowatt-hour be if the city had instead built three coal-fired plants?
3. Suppose that Sea Shell oil company (SS) is pumping oil at a field off the coast of Nigeria. At this site, it has an extraction cost of $30 per barrel for the first 10 million barrels it pumps each year and then $60 per barrel for all subsequent barrels that it pumps each year, up to the site's maximum capacity of 90 million barrels per year. **LO17.4**
   a. Suppose the user cost is $50 per barrel for all barrels and that the current market price for oil is $90 per barrel. How many barrels will SS pump this year? What is the total accounting profit on the total amount of oil it pumps? What is the total economic profit on those barrels of oil?
   b. What if the current market price for oil rises to $120 per barrel, while the user cost remains at $50 per barrel? How many barrels will SS pump and what will be its accounting profit and its economic profit?
   c. If the current market price remains at $120 per barrel but the user cost rises to $95 per barrel, how many barrels will SS pump this year and what will be its accounting profit and its economic profit?
4. Eric and Kyle are fishermen with different equipment and, as a result, different costs for catching fish. Eric's costs for catching fish are $1,000 per ton for the first five tons and then $2,500 per ton for any additional tons. Kyle can harvest

fish at a cost of $3,000 for the first 15 tons and then $1,400 for any additional tons. **LO17.5**

a. If society wants 30 tons of fish and for some reason will only allow one of the two guys to do all the fishing, which one should society choose if it wants to minimize the cost of catching those 30 tons of fish? How much will the total cost of catching the fish be? What will the average cost per ton be for the 30 tons?

b. If society wants 30 tons of fish and wants them for the least cost regardless of who catches them, how much should Eric and Kyle each catch? How much will the total cost of catching 30 tons be? What will the average cost per ton be for the 30 tons?

c. Suppose that Eric and Kyle can both sell whatever amount of fish they catch for $3,000 per ton. Also suppose that Eric is initially given ITQs for 30 tons of fish, while Kyle is given ITQs for zero tons of fish. Suppose that Kyle is willing to pay Eric $550 per ton for as many tons of ITQs as Eric is willing to sell to Kyle. How much profit would Eric make if he used all the ITQs himself? What if Eric sold 25 tons' worth of his ITQs to Kyle while using the other 5 tons of ITQs to fish for himself?

d. What price per ton can Kyle offer to pay Eric for his 25 tons of ITQs such that Eric would make exactly as much money from that deal (in which he sells 25 tons' worth of ITQs to Kyle while using the rest to fish for himself) as he would by using all 30 tons of ITQs for himself?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **18**

Public Finance: Expenditures and Taxes

Learning Objectives

LO18.1 Use a circular flow diagram to illustrate how the allocation of resources is affected by government's revenue and expenditure decisions.

LO18.2 Identify the main categories of government spending and the main sources of government revenue.

LO18.3 List the main categories of federal revenue and spending and describe the difference between marginal and average tax rates.

LO18.4 List the main categories of state and local revenue and spending.

LO18.5 Discuss the magnitude and distribution across job categories of government employment at the local, state, and federal levels.

LO18.6 Summarize the different philosophies regarding the distribution of a nation's tax burden.

LO18.7 Explain the principles relating to tax shifting, tax incidence, and the efficiency losses caused by taxes.

LO18.8 Discuss the probable incidence of U.S. taxes and how the distribution of income between rich and poor is affected by government taxes, transfers, and spending.

As discussed in Chapter 2, the U.S. economy relies heavily on the private sector (households and businesses) and the market system to decide what gets produced, how it gets produced, and who gets the output. But the private sector is not the only entity in the decision process. The public sector (federal, state, and local government) also affects these economic decisions.

Government influences what gets produced and how it gets produced through laws that regulate the activities of private firms and also by directly producing certain goods and services, such as national defense and education. As discussed in Chapter 4, many of these government-produced goods and services are *public goods* that the private sector has trouble producing because of free-rider problems. Also, government influences who receives society's output of goods and services through various taxes and through welfare and income-transfer payments that redistribute income from the rich to the poor.

Government-provided goods, services, and transfer payments are funded by taxes, borrowing, and *proprietary income*—the income that governments receive from running government-owned enterprises such as hospitals, utilities, toll roads, and lotteries.

Public finance is the subdiscipline of economics that studies the various ways in which governments raise and expend money. In this chapter we view the economy through the lens of public finance. Our main goal is to understand how taxes and income transfers not only pay for government-produced goods and services but also affect the distribution of income between rich and poor.

Government and the Circular Flow

LO18.1 Use a circular flow diagram to illustrate how the allocation of resources is affected by government's revenue and expenditure decisions.

In Figure 18.1, we integrate government into the circular flow model first shown in Figure 2.2. Here flows (1) through (4) are the same as the corresponding flows in that figure. Flows (1) and (2) show business expenditures for the resources provided by households. These expenditures are costs to businesses but represent wage, rent, interest, and profit income to households. Flows (3) and (4) show household expenditures for the goods and services produced by businesses.

Now consider what happens when we add government. Flows (5) through (8) illustrate that government makes purchases in both product and resource markets. Flows (5) and (6) represent government purchases of such products as paper, computers, and military hardware from private businesses. Flows (7) and (8) represent government purchases of resources. The federal government employs and pays salaries to members of Congress, the armed forces, Justice Department lawyers, meat inspectors, and so on. State and local governments hire and pay teachers, bus drivers, police, and firefighters. The federal government might also lease or purchase land to expand a military base and a city might buy land on which to build a new elementary school.

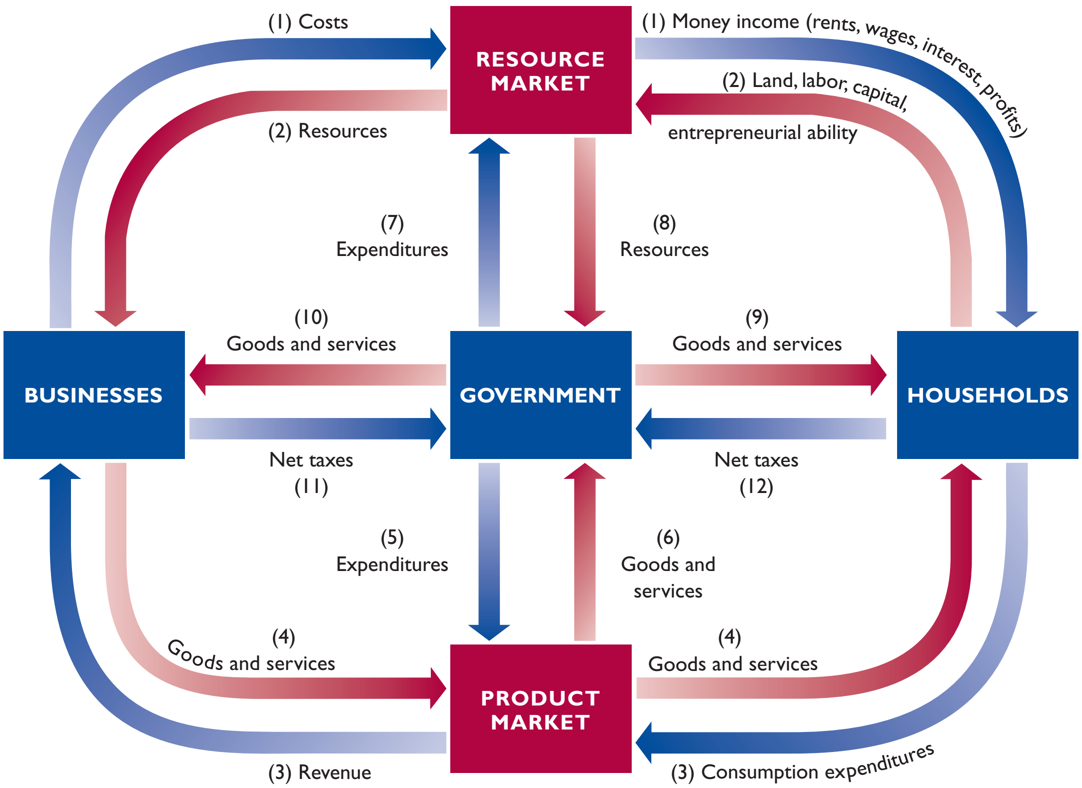
Government then provides goods and services to both households and businesses, as shown by flows (9) and (10). Governments rely on three revenue sources to finance those goods and services: taxes, borrowing, and the proprietary income generated by government-run or government-sponsored businesses like public utilities and state lotteries. These revenues flowing from households and businesses to government are included in flows (11) and (12), which are labeled as "net taxes" for two reasons. First, the vast majority of the money raised by these three revenue sources comes from taxes; thus, it is sensible to have these labels refer to taxes. Second, the labels refer to *net* taxes to indicate that they also include "taxes in reverse" in the form of transfer payments to households and subsidies to businesses. Thus, flow (11) entails various subsidies to farmers, shipbuilders, and airlines as well as income, sales, and excise taxes paid by businesses to government. Most subsidies to business are "concealed" in the form of low-interest loans, loan guarantees, tax concessions, or public facilities provided at prices below their cost. Similarly, flow (12) includes not only taxes (personal income taxes, payroll taxes) collected by government from households but also transfer payments made by government to households. These include welfare payments and Social Security benefits.

Government Finance

LO18.2 Identify the main categories of government spending and the main sources of government revenue.

How large is the U.S. public sector? What are the main expenditure categories of federal, state, and local governments? How are these expenditures financed?

FIGURE 18.1 Government within the circular flow diagram. Government buys products from the product market and employs resources from the resource market to provide goods and services to households and businesses. Government finances its expenditures through the net taxes (taxes minus transfer payments) it receives from households and businesses.



Government Purchases and Transfers

We can get an idea of the size of government's economic role by examining government purchases of goods and services and government transfer payments. There is a significant difference between these two kinds of outlays:

- **Government purchases** are *exhaustive;* the products purchased directly absorb (require the use of) resources and are part of the domestic output. For example, the purchase of a missile absorbs the labor of physicists and engineers along with steel, explosives, and a host of other inputs.
- **Transfer payments** are *nonexhaustive;* they do not directly absorb resources or create output. Social Security benefits, welfare payments, veterans' benefits, and unemployment compensation are examples of transfer payments. Their key characteristic is that recipients make no current contribution to domestic output in return for them.

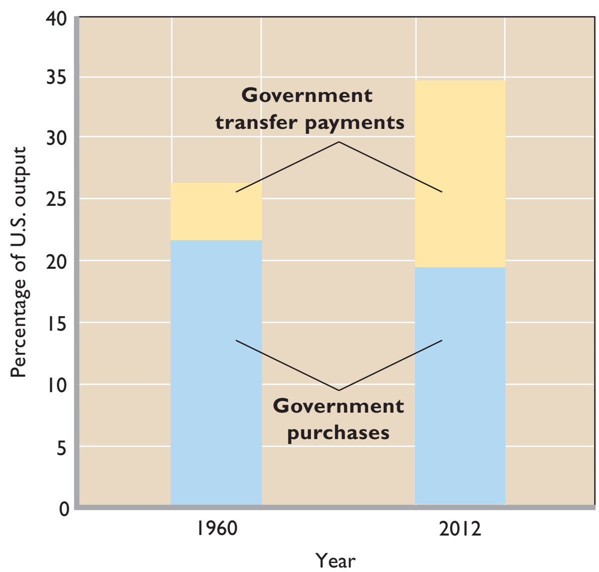
Federal, state, and local governments spent $5,488 billion (roughly $5.5 trillion) in 2012. Of that total, government purchases were $3,083 billion and government transfers were $2,405 billion. Figure 18.2 shows these amounts as percentages of U.S. domestic output for 2012 and compares them to percentages for 1960. Government purchases have declined from about 22 to 20 percent of output since 1960. But transfer payments have tripled as a percentage of output—from 5 percent in 1960 to about 15 percent in 2012. Relative to U.S. output, total government spending is thus higher today than it was 52 years earlier. This means that the tax revenues required to finance government expenditures are also higher. Today, government spending and the tax revenues needed to finance it are about 35 percent of U.S. output.

In 2012 the so-called Tax Freedom Day in the United States was April 17. On that day the average worker had earned enough (from the start of the year) to pay his or her share of the taxes required to finance government spending for the year. Tax Freedom Day arrives even later in several other countries, as indicated in Global Perspective 18.1.

Government Revenues

The funds used to pay for government purchases and transfers come from three sources: taxes, proprietary

FIGURE 18.2 Government purchases, transfers, and total spending as percentages of U.S. output, 1960 and 2012. Government purchases have declined as a percentage of U.S. output since 1960. Transfer payments, however, have increased by more than this drop, raising total government spending (purchases plus transfers) from 27 percent of U.S. GDP in 1960 to about 35 percent today.



Source: Compiled from Bureau of Economic Analysis data, **www.bea.gov**.

income, and funds that are borrowed by selling bonds to the public.

Government Borrowing and Deficit Spending

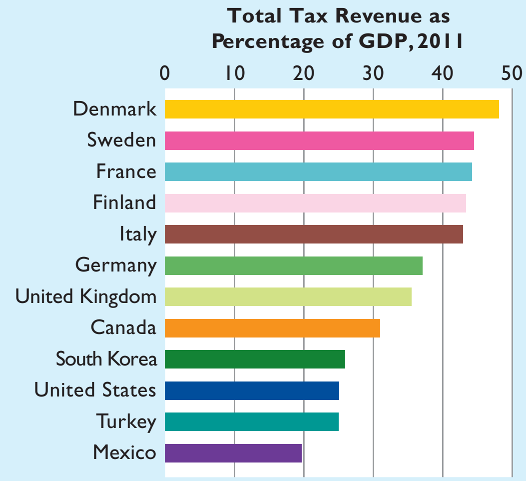
The ability to borrow allows a government to spend more in a given time period than it collects in tax revenues and proprietary income during that period. This flexibility is useful during an economic downturn because a government can use borrowed funds to maintain high levels of spending on goods, services, and transfer payments even if tax revenues and proprietary income are falling due to the slowing economy.

Any money borrowed by a government, however, is money that cannot be put to other uses. During an economic downturn, this opportunity cost is likely to be small because any funds that the government does not borrow are likely to sit idle and unused by other parties due to the lack of economic activity during the downturn. But if the government borrows when the economy is doing well, many economists worry that the opportunity cost may be high. In particular, the government's borrowing may "crowd out" private-sector investment. As an example, a billion dollars borrowed and spent by the federal government on roads is a billion dollars that was not lent to private companies to fund the expansion of factories or the development of new technologies.

GLOBAL PERSPECTIVE 18.1

Total Tax Revenue as a Percentage of Total Output, Selected Nations, 2011*

A nation's "tax burden" is its tax revenue from all levels of government as a percentage of its total output (GDP). Among the world's industrialized nations, the United States has a very moderate tax burden.



*Includes government nontax revenue from fees, charges, fines and sales of government property.
Source: OECD iLibrary, *Taxation: Key Tables from OECD*, **www.oecd-ilibrary.org**, accessed January 18, 2013.

Government spending that is financed by borrowing is often referred to as *deficit spending* because a government's budget is said to be "in deficit" if the government's spending in a given time period exceeds the money that it collects from taxes and proprietary income during that period.

QUICK REVIEW 18.1

- A circular flow diagram can be used to illustrate how the government affects the allocation of resources in the economy through its revenue and expenditure decisions.
- As percentages of GDP, government purchases are 20 percent; government transfers, 15 percent; and the two combined are 35 percent.
- The funds used to pay for government purchases and transfers come from taxes, proprietary income, and borrowing.
- The ability to borrow allows a government to maintain a high level of spending during an economic downturn even if taxes and proprietary income are falling.

Federal Finance

LO18.3 List the main categories of federal revenue and spending and describe the difference between marginal and average tax rates.

Now let's look separately at each of the federal, state, and local units of government in the United States and compare their expenditures and taxes. Figure 18.3 tells the story for the federal government.

Federal Expenditures

Four areas of federal spending stand out: (1) pensions and income security, (2) national defense, (3) health, and (4) interest on the public debt. The *pensions and income security* category includes the many income-maintenance programs for the aged, persons with disabilities or handicaps, the unemployed, the retired, and families with no breadwinner. This category—dominated by the $597 billion pension portion of the Social Security program—accounts for 37 percent of total federal expenditures. *National defense* accounts for about 19 percent of the federal budget, underscoring the high cost of military preparedness. *Health* reflects the cost of government health programs for the retired (Medicare) and poor (Medicaid). *Interest on the public debt* accounts for 6 percent of federal spending.

Federal Tax Revenues

The revenue side of Figure 18.3 shows that the personal income tax, payroll taxes, and the corporate income tax are the largest revenue sources, accounting respectively for 46, 35, and 10 cents of each dollar collected.

Personal Income Tax The **personal income tax** is the kingpin of the federal tax system and merits special comment. This tax is levied on *taxable income*, that is, on the incomes of households and unincorporated businesses after certain exemptions ($3,650 for each household member) and deductions (business expenses, charitable contributions, home mortgage interest payments, certain state and local taxes) are taken into account.

The federal personal income tax is a *progressive tax*, meaning that people with higher incomes pay a larger percentage of their incomes as taxes than do people with lower incomes. The progressivity is achieved by applying higher tax rates to successive layers or brackets of income.

Columns 1 and 2 in Table 18.1 show the mechanics of the income tax for a married couple filing a joint return in 2012. Note that a 10 percent tax rate applies to all taxable income up to $17,850 and a 15 percent rate applies to additional income up to $72,500. The rates on additional layers of income then go up to 25, 28, 33, and 39.6 percent.

The tax rates shown in column 2 in Table 18.1 are marginal tax rates. A **marginal tax rate** is the rate at which the tax is paid on each *additional* unit of taxable income. Thus, if a couple's taxable income is $80,000, they will pay the marginal rate of 10 percent on each dollar from $1 to $17,850, 15 percent on each dollar from $17,851 to $72,500, and 25 percent on each dollar from $72,501 to $80,000. You should confirm that their total income tax is $13,643.

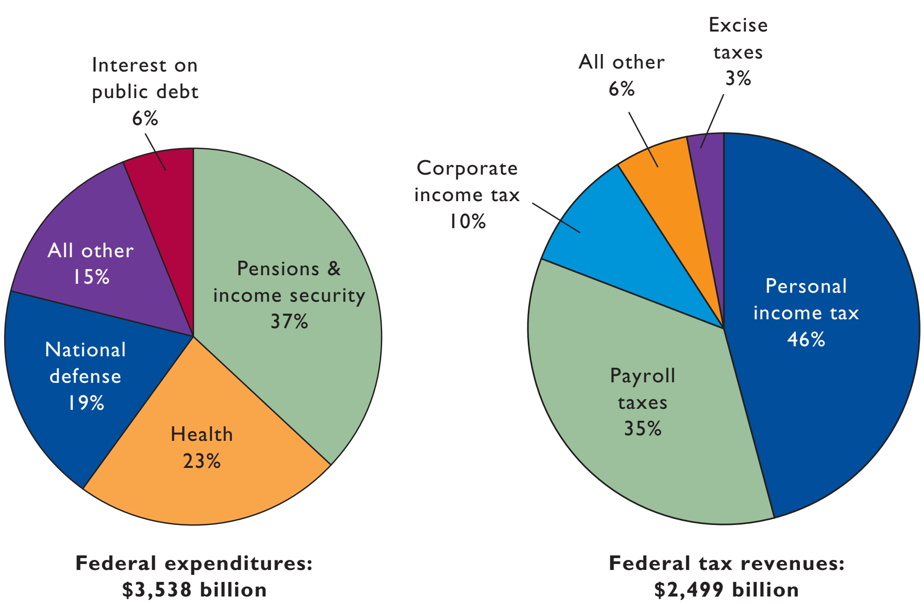


FIGURE 18.3 Federal expenditures and tax revenues, 2012. Federal expenditures are dominated by spending for pensions and income security, health, and national defense. A full 81 percent of federal tax revenue is derived from just two sources: the personal income tax and payroll taxes. The $1,039 billion difference between expenditures and revenues reflects a budget deficit.

Source: U.S. Treasury, *Combined Statement of Receipts, Outlays, and Balances, 2012,* **fms.treas.gov**.

TABLE 18.1 **Federal Personal Income Tax Rates, 2013***

(1) Total Taxable Income	(2) Marginal Tax Rate,%	(3) Total Tax on Highest Income In Bracket	(4) Average Tax Rate on Highest Income in Bracket,% (3) ÷ (1)
$0–$17,850	10	$ 1,785	10
$17,851–$72,500	15	9,983	14
$72,501–$146,400	25	28,458	19
$146,401–$223,050	28	49,920	22
$223,051–$398,350	33	107,769	27
$398,351–$450,000	35	125,847	28
$450,001 and above	39.6		

*For a married couple filing a joint return.

The marginal tax rates in column 2 overstate the personal income tax bite because the rising rates in that column apply only to the income within each successive tax bracket. To get a better idea of the tax burden, we must consider average tax rates. The **average tax rate** is the total tax paid divided by total taxable income. The couple in our previous example is in the 25 percent tax bracket because they pay a top marginal tax rate of 25 percent on the highest dollar of their income. But their *average* tax rate is 17 percent (= $13,643/$80,000).

WORKED PROBLEMS

W18.1
Taxes and progressivity

As we will discuss in more detail shortly, a tax whose average rate rises as income increases is said to be a *progressive tax* because it claims both a progressively larger absolute amount of income as well as a progressively larger proportion of income as income rises. Thus we can say that the federal personal income tax is progressive.

Payroll Taxes Social Security contributions are **payroll taxes**—taxes based on wages and salaries—used to finance two compulsory federal programs for retired workers: Social Security (an income-enhancement program) and Medicare (which pays for medical services). Employers and employees pay these taxes equally. In 2012, employees and employers each paid 7.65 percent on the first $110,100 of an employee's annual earnings and 1.45 percent on all additional earnings.

Corporate Income Tax The federal government also taxes corporate income. The **corporate income tax** is levied on a corporation's profit—the difference between its total revenue and its total expenses. For almost all corporations, the tax rate is 35 percent.

Excise Taxes Taxes on commodities or on purchases take the form of **sales and excise taxes.** The two differ primarily in terms of coverage. Sales taxes fall on a wide range of products, whereas excises are levied individually on a small, select list of commodities. An additional difference is that sales taxes are calculated as a percentage of the price paid for a product, whereas excise taxes are levied on a per-unit basis—for example, $2 per pack of cigarettes or $0.50 per gallon of gasoline.

As Figure 18.3 suggests, the federal government collects excise taxes of various rates (on the sale of such commodities as alcoholic beverages, tobacco, and gasoline) but does not levy a general sales tax; sales taxes are, however, the primary revenue source of most state governments.

State and Local Finance

LO18.4 List the main categories of state and local revenue and spending.

State and local governments have different mixes of revenues and expenditures than the federal government has.

State Finances

Figure 18.4 shows that the primary source of tax revenue for state governments is sales and excise taxes, which account for about 49 percent of all their tax revenue. State personal income taxes, which have much lower rates than the federal income tax, are the second most important source of state tax revenue. They bring in about 34 percent of total state tax revenue. Corporate income taxes and license fees account for most of the remainder of state tax revenue.

Education expenditures account for about 36 percent of all state spending. State expenditures on public welfare are next in relative weight, at about 30 percent of the total. States also spend heavily on health and hospitals (8 percent), highway maintenance and construction (7 percent), and public safety (4 percent). That leaves about 15 percent of all state spending for a variety of other purposes.

These tax and expenditure percentages combine data from all the states, so they reveal little about the finances of individual states. States vary significantly in the taxes levied. Thus, although personal income taxes are a major source of revenue for all state governments combined, seven states do not levy a personal income tax. Also, there are great variations in the sizes of tax revenues and disbursements among the states, both in the aggregate and as percentages of personal income.

Forty-three states augment their tax revenues with state-run lotteries to help close the gap between their tax

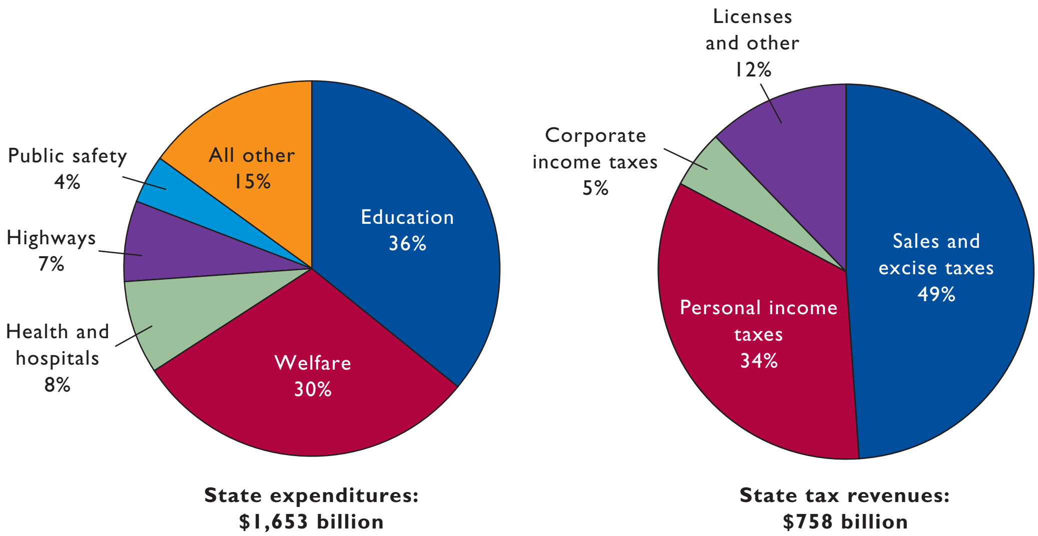


FIGURE 18.4 State expenditures and tax revenues, 2011. State governments spend largely on education and welfare. Their primary source of tax revenue is sales and excise taxes. The deficit between state expenditures and state tax revenues is filled by proprietary income and intergovernmental grants from the federal government. The state expenditures numbers here include state grants to local governments.

Source: U.S. Census Bureau, *2011 Annual Survey of State Government Finances,* **www.census.gov**.

receipts and expenditures. Individual states also receive large intergovernmental grants from the federal government. In fact, about 24 percent of their total revenue is in that form. States also take in revenue from miscellaneous sources such as state-owned utilities and liquor stores.

Local Finances

The local levels of government include counties, municipalities, townships, and school districts as well as cities and towns. Figure 18.5 shows that local governments obtain about 75 percent of their tax revenue from **property taxes.** Sales and excise taxes contribute about 16 percent of all local government tax revenue.

About 44 percent of local government expenditures go to education. Welfare, health, and hospitals (12 percent); public safety (11 percent); housing, parks, and sewerage (11 percent); and streets and highways (6 percent) are also major spending categories.

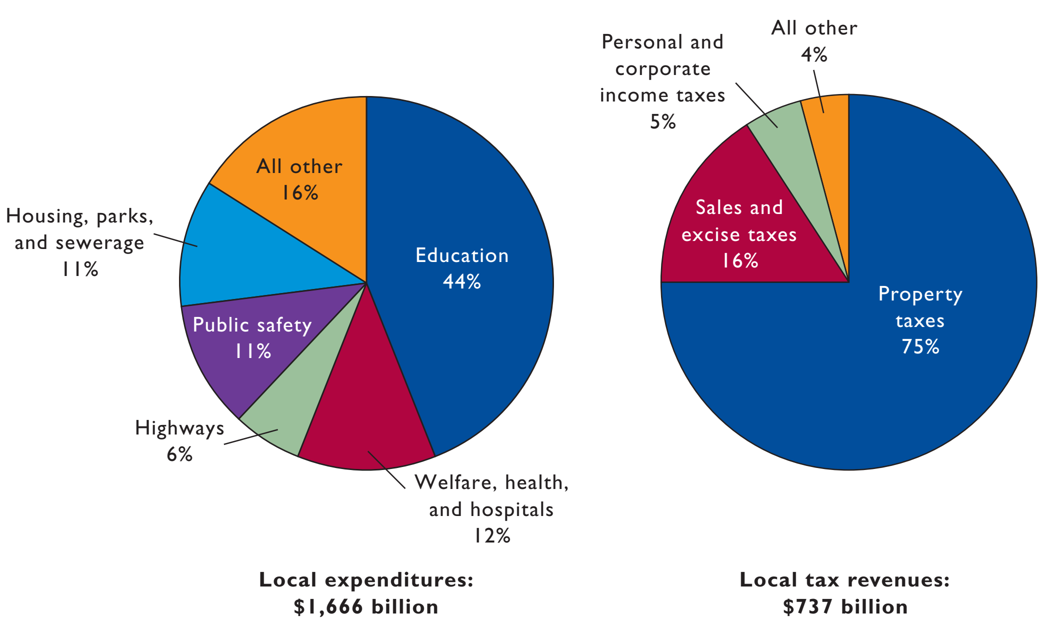


FIGURE 18.5 Local expenditures and tax revenues, 2010. The expenditures of local governments go largely to education, while a large majority of local tax collections are obtained via property taxes. The large deficit between local expenditures and local tax revenues is filled by proprietary income and federal and state intergovernmental grants.

Source: U.S. Census Bureau, *2010 Annual Surveys of State and Local Government Finances,* **www.census.gov**.

CONSIDER THIS . . .

State Lotteries: A Good Bet?

State lotteries generated about \$54.7 billion in revenue in 2011. Of that amount, \$33.8 billion went to prizes and \$2.8 billion went to administrative costs. That left \$18.3 billion that could be spent by the states as they saw fit.

Though nowadays common, state lotteries are still controversial. Critics argue that (1) it is morally wrong for states to sponsor gambling; (2) lotteries generate compulsive gamblers who impoverish themselves and their families; (3) low-income families spend a larger portion of their incomes on lotteries than do high-income families; (4) as a cash business, lotteries attract criminals and other undesirables; and (5) lotteries send the message that luck and fate—rather than education, hard work, and saving—are the route to wealth.

Defenders contend that (1) lotteries are preferable to taxes because they are voluntary rather than compulsory; (2) they are a relatively painless way to finance government services such as education, medical care, and welfare; and (3) lotteries compete with illegal gambling and are thus socially beneficial in curtailing organized crime.

As a further point for debate, also note that state lotteries are monopolies, with states banning competing private lotteries. The resulting lack of competition allows many states to restrict prizes to only about half the money wagered. These payout rates are substantially lower than the 80–95 percent payout rates typically found in private betting operations such as casinos.

Thus, while lotteries are indeed voluntary, they are overpriced and underprovided relative to what would happen if there were a free market in lotteries. But, then again, a free market in lotteries would eliminate monopoly profits for state lotteries and possibly add government costs for regulation and oversight. Consequently, the alternative of allowing a free market in lottery tickets and then taxing the firms selling lottery tickets would probably net very little additional revenue to support state spending programs.

The tax revenues of local government cover less than one-half of their expenditures. The bulk of the remaining revenue comes from intergovernmental grants from the federal and state governments. Also, local governments receive considerable amounts of proprietary income, for example, revenue from government-owned utilities providing water, electricity, natural gas, and transportation.

Local, State, and Federal Employment

LO18.5 Discuss the magnitude and distribution across job categories of government employment at the local, state, and federal levels.

In 2011, U.S. governments (local, state, and federal) employed about 21.9 million workers, or about 16 percent of the U.S. labor force. Figure 18.6 shows the percentages of these government employees assigned to different tasks at both the federal level and the state and local level.

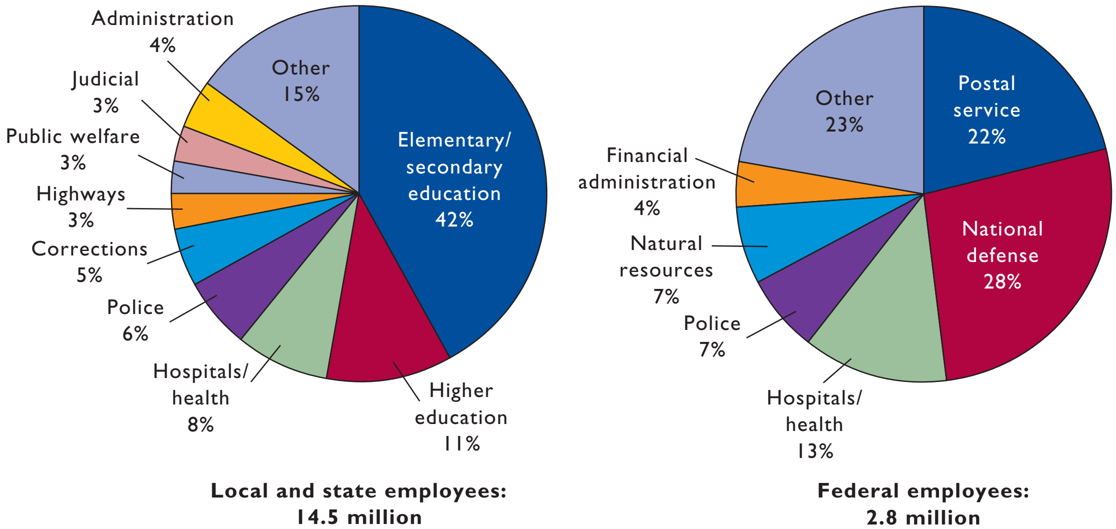
As Figure 18.6 makes clear, the types of jobs done by government workers depend on the level of government. Over half of state and local government employment is focused on education. The next largest sector is hospitals and health care, which accounts for about 8 percent of state and local government employment. Police and corrections make up another 11 percent. Smaller categories like highways, public welfare, and judicial together combine for less than 10 percent of state and local employment. The "other" category includes workers in areas such as parks and recreation, fire fighting, transit, and libraries.

Half of federal government jobs are in national defense or the postal service. A further 12 percent of federal government jobs are in hospitals or health care. The natural resources, police, and financial administration categories each accounts for between 4 and 7 percent of federal employment. The "other" category at the federal level is composed of workers in areas such as justice and law, corrections, air transportation, and social insurance administration.

QUICK REVIEW 18.2

- Income security and national defense are the main categories of federal spending; personal income, payroll, and corporate income taxes are the primary sources of federal revenue.
- States rely on sales and excise taxes for revenue; their spending is largely for education and public welfare.
- Education is the main expenditure for local governments, most of whose revenue comes from property taxes.
- State and local employment is dominated by education, while federal employment is dominated by national defense and the postal service.

FIGURE 18.6 Job functions of state and local employees and federal employees, 2011. A majority of state and local workers are employed in education. Federal employment is dominated by the postal service and national defense, which together employ just over half of federal employees.



Source: U.S. Census Bureau, *State and Local Government Employment and Payroll Data, by State and Function,* and *Federal Government Employment by Function,* **www.census.gov**.

Apporitioning the Tax Burden

LO18.6 Summarize the different philosophies regarding the distribution of a nation's tax burden.

Taxes are the major source of funding for the goods and services provided by government and the wages and salaries paid to government workers. Without taxes, there would be no public schools, no national defense, no public highways, no courts, no police, and no other government-provided public and quasi-public goods. As stated by Supreme Court Justice Oliver Wendell Holmes, "Taxes are the price we pay for civilization."

But taxes are controversial. To begin with, many people would prefer to obtain government goods and services without paying for them. Many others argue that certain taxes cause more harm than good, either by discouraging beneficial economic activity or by unfairly reducing the income flowing to workers and investors. And millions more chafe at the huge variety of taxes that governments levy, including income taxes, Social Security taxes, Medicare taxes, property taxes, sales taxes, liquor taxes, cigarette taxes, cell phone taxes, hotel taxes, gasoline taxes, profit taxes, and estate taxes. As the Beatles put it in their song "Taxman": "If you drive a car, I'll tax the street. If you try to sit, I'll tax your seat."

For these and other reasons, people are intently focused on the overall level of taxes, the amount they must personally pay, and the idea of tax fairness (which is often defined in terms of their own circumstances).

The public's attention to taxes has spurred public finance economists to undertake considerable research into the size, distribution, and impact of the total costs that taxes impose on society—the so-called tax burden. Their investigations reveal with reasonable clarity both the size of the tax burden as well as how it is apportioned across the income distribution.

Whether you consider their findings to be good news or bad news, however, depends significantly on your opinion about the fairest way to allocate taxes and the tax burden. So before turning to their findings, let's first discuss some of the major philosophical viewpoints regarding taxation.

Benefits Received versus Ability to Pay

Two basic philosophies coexist on how the economy's tax burden should be apportioned.

Benefits-Received Principle The **benefits-received principle** of taxation asserts that households should purchase the goods and services of government in the same way they buy other commodities. Those who benefit most from government-supplied goods or services should pay the taxes necessary to finance them. A few public goods are now financed on this basis. For example, money collected as gasoline taxes is typically used to finance highway construction and repairs. Thus people who benefit from good roads pay the cost of those roads. Difficulties immediately arise, however, when we consider widespread application of the benefits-received principle:

- How will the government determine the benefits that individual households and businesses receive from national defense, education, the court system, and police and fire protection? Recall from Chapter 4 that public goods are characterized by nonrivalry and nonexcludability. So benefits from public goods are especially widespread and diffuse. Even in the seemingly straightforward case of highway financing it is difficult to measure benefits. Good roads benefit owners of cars in different degrees. But others also benefit. For example, businesses benefit because good roads bring them workers and customers.
- The benefits-received principle cannot logically be applied to income redistribution programs. It would be absurd and self-defeating to ask poor families to pay the taxes needed to finance their welfare payments. It would also be self-defeating to tax only unemployed workers to finance the unemployment benefits they receive.

Ability-to-Pay Principle The **ability-to-pay principle** of taxation asserts that the tax burden should be apportioned according to taxpayers' income and wealth. In practice, this means that individuals and businesses with larger incomes should pay more taxes in both absolute and relative terms than those with smaller incomes.

In justifying the ability-to-pay principle, proponents contend that each additional dollar of income received by a household yields a smaller amount of satisfaction or marginal utility when it is spent. Because consumers act rationally, the first dollars of income received in any time period will be spent on high-urgency goods that yield the greatest marginal utility. Successive dollars of income will go for less urgently needed goods and finally for trivial goods and services. This means that a dollar taken through taxes from a poor person who has few dollars represents a greater utility sacrifice than a dollar taken through taxes from a rich person who has many dollars. To balance the sacrifices that taxes impose on income receivers, taxes should be apportioned according to the amount of income a taxpayer receives.

This argument is appealing, but application problems arise here too. Although we might agree that the household earning $100,000 per year has a greater ability to pay taxes than a household receiving $10,000, we don't know exactly how much more ability to pay the first family has. Should the wealthier family pay the *same* percentage of its larger income, and hence a larger absolute amount, as taxes? Or should it be made to pay a *larger* percentage of its income as taxes? And how much larger should that percentage be? Who is to decide?

There is no scientific way of making utility comparisons among individuals and thus of measuring someone's relative ability to pay taxes. That is the main problem. In practice, the solution hinges on guesswork, the tax views of the political party in power, expediency, and how urgently the government needs revenue.

Progressive, Proportional, and Regressive Taxes

Any discussion of taxation leads ultimately to the question of tax rates. Taxes are classified as progressive, proportional, or regressive, depending on the relationship between average tax rates and taxpayer incomes. We focus on incomes because all taxes—whether on income, a product, a building, or a parcel of land—are ultimately paid out of someone's income.

- A tax is **progressive** if its average rate increases as income increases. Such a tax claims not only a larger absolute (dollar) amount but also a larger percentage of income as income increases.
- A tax is **regressive** if its average rate declines as income increases. Such a tax takes a smaller proportion of income as income increases. A regressive tax may or may not take a larger absolute amount of income as income increases. (You may want to develop an example to substantiate this fact.)
- A tax is **proportional** if its average rate *remains the same* regardless of the size of income. Proportional income taxes are often referred to as *flat taxes* or *flat-rate taxes* because their average rates do not vary with (are flat with respect to) income levels.

We can illustrate these ideas with the personal income tax. Suppose tax rates are such that a household pays 10 percent of its income in taxes regardless of the size of its income. This is a *proportional* income tax. Now suppose the rate structure is such that a household with an annual taxable income of less than $10,000 pays 5 percent in income taxes; a household with an income of $10,000 to $20,000 pays 10 percent; one with a $20,000 to $30,000 income pays 15 percent; and so forth. This is a *progressive* income tax. Finally, suppose the rate declines as taxable income rises: You pay 15 percent if you earn less than $10,000; 10 percent if you earn $10,000 to $20,000; 5 percent if you earn $20,000 to $30,000; and so forth. This is a *regressive* income tax.

In general, progressive taxes are those that fall relatively more heavily on people with high incomes; regressive taxes are those that fall relatively more heavily on the poor.

Applications Let's examine the progressivity, or regressivity, of several taxes.

Personal Income Tax As noted earlier, the federal personal income tax is progressive, with marginal tax rates (those assessed on additional income) ranging from 10 to 39.6 percent in 2013. Rules that allow individuals to deduct from income interest on home mortgages and property taxes and that exempt interest on state and local bonds from taxation tend to make the tax less progressive than these marginal rates suggest. Nevertheless, average tax rates rise with income.

Sales Taxes At first thought, a general sales tax with, for example, a 5 percent rate would seem to be proportional. But in fact it is regressive with respect to income. A larger portion of a low-income person's income is exposed to the tax than is the case for a high-income person; the rich pay no tax on the part of income that is saved, whereas the poor are unable to save. Example: "Low-income" Smith has an income of $15,000 and spends it all. "High-income" Jones has an income of $300,000 but spends only $200,000 and saves the rest. Assuming a 5 percent sales tax applies to all expenditures of each individual, we find that Smith pays $750 (5 percent of $15,000) in sales taxes and Jones pays $10,000 (5 percent of $200,000). But Smith pays $750/$15,000, or 5 percent of income as sales taxes while Jones pays $10,000/$300,000, or 3.3 percent of income. The general sales tax therefore is regressive.

Corporate Income Tax The federal corporate income tax is essentially a proportional tax with a flat 35 percent tax rate. In the short run, the corporate owners (shareholders) bear the tax through lower dividends and share values. In the long run, workers may bear some of the tax since it reduces the return on investment and therefore slows capital accumulation. It also causes corporations to relocate to other countries that have lower tax rates. With less capital per worker, U.S. labor productivity may decline and wages may fall. To the extent this happens, the corporate income tax may be somewhat regressive.

Payroll Taxes Payroll taxes are taxes levied upon wages and salaries by certain states as well as by the federal government. The federal payroll tax is known as the FICA tax after the Federal Insurance Contributions Act, which mandated one payroll tax to fund the Social Security program and another to fund the Medicare program.

Both taxes are split equally between employer and employee. Thus, the 12.4 percent Social Security tax is split in half, with 6.2 percent paid by employees and an additional 6.2 percent paid by employers. In the same way, the 2.9 percent Medicare tax is also split in half, with 1.45 percent paid by employees and 1.45 percent paid by employers.

Crucially, however, only the Medicare tax applies to all wage and salary income without limit. The Social Security

CONSIDER THIS . . .

The VAT: A Very Alluring Tax?

A value-added tax (VAT) is like a retail sales tax except that it applies only to the *difference* between the value of a firm's sales and the value of its purchases from other firms. For instance, Intel would pay the VAT—say, 7 percent—only on the difference between the value of the microchips it sells and the value of the materials used to make them. Dell, Lenovo, and other firms that buy chips and other components to make computers would subtract the value of their materials from the value of their sales of personal computers. They would pay the 7 percent tax on that difference—on the value that *they* added.

Economists reason that because the VAT would apply to all firms, sellers could shift their VATs to buyers in the form of higher prices without having to worry that their higher prices might cause them to lose sales to competitors. Final consumers, who cannot shift the tax, would be the ones who ultimately end up paying the full VAT as 7 percent higher prices. So the VAT would amount to a national sales tax on consumer goods.

Most other nations besides the United States have a VAT in addition to other taxes. Why the attraction? Proponents argue that it encourages savings and investment because it penalizes consumption. Unlike income taxes and profits taxes, which reduce the returns to working and investing, the VAT only taxes consumption. Thus, people might be expected to save and invest more if the government switched from taxing income and profits to taxing consumption via a VAT.

Opponents counter, however, that the VAT discourages savings and investment just as much as do income and profit taxes because the whole point of working hard, saving, and investing is the ability to reward yourself in the future with increased consumption. By making consumption more expensive, the VAT reduces this future reward. Also, because VATs are regressive, opponents argue that VATs lead to higher and more progressive income taxes as governments try to use the progressivity of income taxes to counter the regressivity of the VAT. Finally, critics note that the VAT is deeply buried within product prices and therefore is a *hidden tax*. Such taxes are usually easier to increase than other taxes and therefore may result in excessively large government.

tax, by contrast, is "capped," meaning that it applies only up to a certain limit, or cap. In 2012, the cap was $110,100.

The fact that the Social Security tax applies only on income below the cap implies that the FICA tax is regressive. To see this, consider a person with $110,100 in wage income. He would pay $8,422.65, or 7.65 percent (= 6.2 percent + 1.45 percent) of his wages in FICA taxes. By contrast, someone with twice that income, or $220,200, would pay $10,019.10 (= $8,422.65, on the first $110,100 + $1,596.45 on the second $110,100), which is only 4.6 percent of his wage income. Thus the average FICA tax falls as income rises, thereby confirming that the FICA tax is regressive.

But payroll taxes are even more regressive than suggested by this example because they only apply to wage and salary income. People earning high incomes tend to derive a higher percentage of their total incomes from nonwage sources like rents and dividends than do people who have incomes below the $110,100 cap on which Social Security taxes are paid. Thus, if our individual with the $220,200 of wage income also received $220,200 of nonwage income, his $10,019.10 of FICA tax would be only 2.3 percent of his total income of $440,400.

Property Taxes Most economists conclude that property taxes on buildings are regressive for the same reasons as are sales taxes. First, property owners add the tax to the rents that tenants are charged. Second, property taxes, as a percentage of income, are higher for low-income families than for high-income families because the poor must spend a larger proportion of their incomes for housing.

Tax Incidence and Efficiency Loss

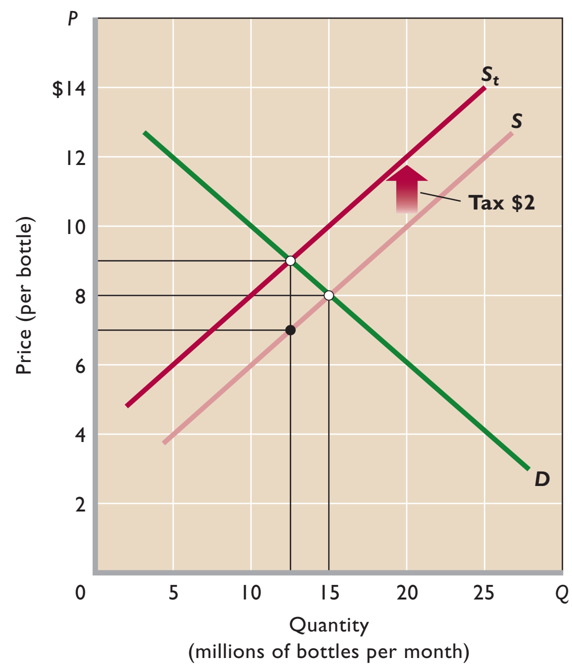
LO18.7 Explain the principles relating to tax shifting, tax incidence, and the efficiency losses caused by taxes.

Determining whether a particular tax is progressive, proportional, or regressive is complicated because those on whom taxes are levied do not always pay the taxes. This is true because some or all of the value of the tax may be passed on to others. We therefore need an understanding of **tax incidence,** the degree to which a tax falls on a particular person or group. The tools of elasticity of supply and demand will help. Let's focus on a hypothetical excise tax levied on wine producers. Do the producers really pay this tax, or is some fraction of the tax shifted to wine consumers?

Elasticity and Tax Incidence

In Figure 18.7, *S* and *D* represent the pretax market for a certain domestic wine; the no-tax equilibrium price and quantity are $8 per bottle and 15 million bottles. Suppose

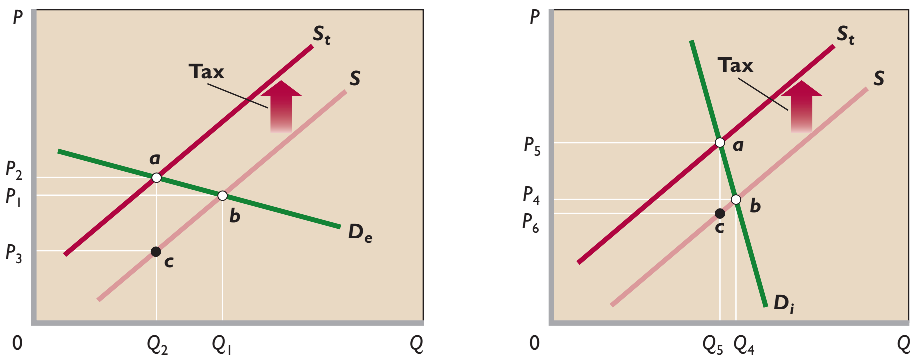
FIGURE 18.7 The incidence of an excise tax. An excise tax of a specified amount (here, $2 per unit) shifts the supply curve upward by the amount of the tax per unit: the vertical distance between *S* and S_t. This results in a higher price (here, $9) to consumers and a lower after-tax price (here, $7) to producers. Thus consumers and producers share the burden of the tax in some proportion (here, equally at $1 per unit).



that government levies an excise tax of $2 per bottle at the winery. Who will actually pay this tax?

Division of Burden Since the government imposes the tax on the sellers (suppliers), we can view the tax as an addition to the marginal cost of the product. Now sellers must get $2 more for each bottle to receive the same per-unit profit they were getting before the tax. While sellers are willing to offer, for example, 5 million bottles of untaxed wine at $4 per bottle, they must now receive $6 per bottle (= $4 + $2 tax) to offer the same 5 million bottles. The tax shifts the supply curve upward (leftward) as shown in Figure 18.7, where S_t is the "after-tax" supply curve.

The after-tax equilibrium price is $9 per bottle, whereas the before-tax equilibrium price was $8. So, in this case, consumers pay half the $2 tax as a higher price; producers pay the other half in the form of a lower after-tax per-unit revenue. That is, after remitting the $2 tax per unit to government, producers receive $7 per bottle, or $1 less than the $8 before-tax price. So, in this case, consumers and producers share the burden of the tax equally: Half of the $2 per bottle tax is shifted to consumers in the form of a higher price and half is paid by producers.



(a) Tax incidence with elastic demand

(b) Tax incidence with inelastic demand

FIGURE 18.8 Demand elasticity and the incidence of an excise tax. (a) If demand is elastic in the relevant price range, price rises modestly (P_1 to P_2) when an excise tax is levied. Hence, the producers bear most of the tax burden. (b) If demand is inelastic, the price increases substantially (P_4 to P_5) and most of the tax is borne by consumers.

Note also that the equilibrium quantity declines because of the tax levy and the higher price that it imposes on consumers. In Figure 18.7 that decline in quantity is from 15 million bottles to 12.5 million bottles per month.

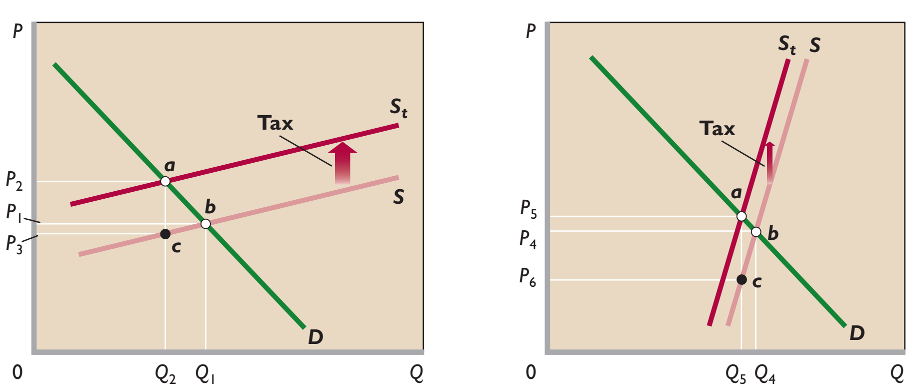
Elasticities If the elasticities of demand and supply were different from those shown in Figure 18.7, the incidence of tax would also be different. Two generalizations are relevant.

With a specific supply, the more inelastic the demand for the product, the larger is the portion of the tax shifted to consumers. To verify this, sketch graphically the extreme cases in which demand is perfectly elastic and perfectly inelastic. In the first case, the incidence of the tax is entirely on sellers; in the second, the tax is shifted entirely to consumers.

Figure 18.8 contrasts the more usual cases where demand is either relatively elastic or relatively inelastic in the relevant price range. With elastic demand (Figure 18.8a), a small portion of the tax ($P_2 - P_1$) is shifted to consumers and most of the tax ($P_1 - P_3$) is borne by the producers. With inelastic demand (Figure 18.8b), most of the tax ($P_5 - P_4$) is shifted to consumers and only a small amount ($P_4 - P_6$) is paid by producers. In both graphs the per-unit tax is represented by the vertical distance between S_t and S.

Note also that the decline in equilibrium quantity (from Q_1 to Q_2 in Figure 18.8a and from Q_4 to Q_5 in Figure 18.8b) is smaller when demand is more inelastic. This is the basis of our previous applications of the elasticity concept to taxation in earlier chapters: Revenue-seeking legislatures place heavy excise taxes on liquor, cigarettes, automobile tires, telephone service, and other products whose demand is thought to be inelastic. Since demand for these products is relatively inelastic, the tax does not reduce sales by much, so the tax revenue stays high.

The second generalization is that, with a specific demand, the more inelastic the supply, the larger is the portion of the tax borne by producers. When supply is elastic (Figure 18.9a), consumers bear most of the tax ($P_2 - P_1$)



(a) Tax incidence with elastic supply

(b) Tax incidence with inelastic supply

FIGURE 18.9 Supply elasticity and the incidence of an excise tax. (a) With elastic supply, an excise tax results in a large price increase (P_1 to P_2) and the tax is therefore paid mainly by consumers. (b) If supply is inelastic, the price rise is small (P_4 to P_5) and sellers bear most of the tax.

while producers bear only a small portion ($P_1 - P_3$) themselves. But where supply is inelastic (Figure 18.9b), the reverse is true: The major portion of the tax ($P_4 - P_6$) falls on sellers, and a relatively small amount ($P_5 - P_4$) is shifted to buyers. The equilibrium quantity also declines less with an inelastic supply than it does with an elastic supply.

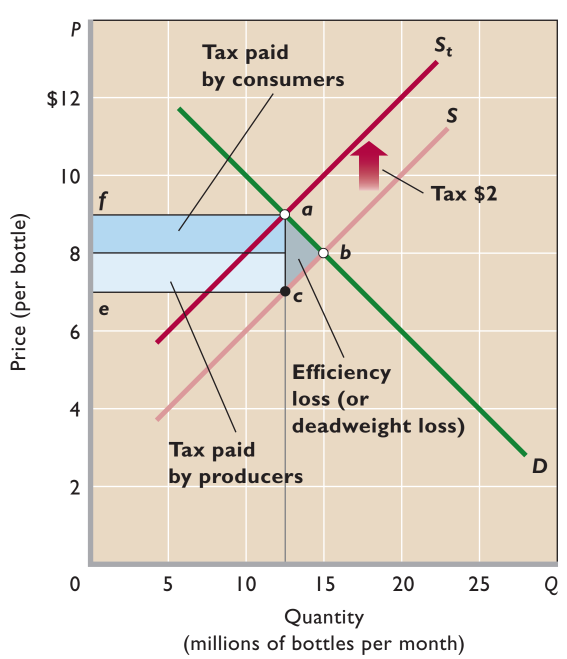
Gold is an example of a product with an inelastic supply and therefore one where the burden of an excise tax (such as an extraction tax) would mainly fall on producers. On the other hand, because the supply of baseballs is relatively elastic, producers would pass on to consumers much of an excise tax on baseballs.

Efficiency Loss of a Tax

We just observed that producers and consumers typically each bears part of an excise tax levied on producers. Let's now look more closely at the overall economic effect of the excise tax. Consider Figure 18.10, which is identical to Figure 18.7 but contains the additional detail we need for our discussion.

Tax Revenues In our example, a \$2 excise tax on wine increases its market price from \$8 to \$9 per bottle and reduces the equilibrium quantity from 15 million bottles to 12.5 million. Government tax revenue is \$25 million (= \$2 × 12.5 million bottles), an amount shown as the rectangle *efac* in Figure 18.10. The elasticities of supply and demand in this case are such that consumers and producers each pays half this total amount, or \$12.5 million apiece (= \$1 × 12.5 million bottles). The government uses this \$25 million of tax revenue to provide public goods and services. So this transfer of dollars from consumers and producers to government involves no loss of well-being to society.

FIGURE 18.10 Efficiency loss (or deadweight loss) of a tax. The levy of a \$2 tax per bottle of wine increases the price per bottle from \$8 to \$9 and reduces the equilibrium quantity from 15 million to 12.5 million. Tax revenue to the government is \$25 million (area *efac*). The efficiency loss of the tax arises from the 2.5 million decline in output; the amount of that loss is shown as triangle *abc*.



Efficiency Loss The \$2 tax on wine does more than require consumers and producers to pay \$25 million of taxes; it also reduces the equilibrium amount of wine produced and consumed by 2.5 million bottles. The fact that consumers and producers demanded and supplied 2.5 million more bottles of wine before the tax means that those 2.5 million bottles provided benefits in excess of their production costs. This is clear from the following analysis.

Segment *ab* of demand curve *D* in Figure 18.10 indicates the willingness to pay—the marginal benefit—associated with each of the 2.5 million bottles consumed before (but not after) the tax. Segment *cb* of supply curve *S* reflects the marginal cost of each of the bottles of wine. For all but the very last one of these 2.5 million bottles, the marginal benefit (shown by a point on *ab*) exceeds the marginal cost (shown by a point on *cb*). Not producing these 2.5 million bottles of wine reduces well-being by an amount represented by the triangle *abc*. The area of this triangle identifies the **efficiency loss of the tax** (also called the *deadweight loss of the tax*). This loss is society's sacrifice of net benefit because the tax reduces production and consumption of the product below their levels of economic efficiency, where marginal benefit and marginal cost are equal.

Role of Elasticities Most taxes create some degree of efficiency loss, but just how much depends on the supply and demand elasticities. Glancing back at Figure 18.8, we see that the efficiency loss area *abc* is greater in Figure 18.8a, where demand is relatively elastic, than in Figure 18.8b, where demand is relatively inelastic. Similarly, area *abc* is greater in Figure 18.9a than in Figure18.9b, indicating a larger efficiency loss where supply is more elastic. Other things equal, the greater the elasticities of supply and demand, the greater the efficiency loss of a particular tax.

Two taxes yielding equal revenues do not necessarily impose equal costs on society. The government must keep this fact in mind in designing a tax system to finance beneficial public goods and services. In general, it should minimize the efficiency loss of the tax system in raising any specific dollar amount of tax revenue.

Qualifications We must acknowledge, however, that other tax goals may be as important as, or even more important than, minimizing efficiency losses from taxes. Here are two examples:

- ***Redistributive goals*** Government may wish to impose progressive taxes as a way to redistribute income. The 10 percent excise tax the federal government placed on selected luxuries in 1990 was an example. Because the demand for luxuries is elastic, substantial efficiency losses from this tax were to be expected. However, Congress apparently concluded that the benefits from the redistribution effects of the tax would exceed the efficiency losses.

  Ironically, in 1993 Congress repealed the luxury taxes on personal airplanes and yachts, mainly because the taxes had reduced quantity demanded so much that widespread layoffs of workers were occurring in those industries. But the 10 percent tax on luxury automobiles remained in place until it expired in 2003.
- ***Reducing negative externalities*** Our analysis of the efficiency loss of a tax assumes no negative externalities arising from either the production or consumption of the product in question. Where such spillover costs occur, an excise tax on producers might actually improve allocative efficiency by reducing output and thus lessening the negative externality. For example, the $2 excise tax on wine in our example might be part of a broader set of excise taxes on alcoholic beverages. The government may have concluded that the consumption of these beverages produces certain negative externalities. Therefore, it might have purposely levied this $2 tax to shift the market supply curve in Figure 18.10 to increase the price of wine, decrease alcohol consumption, and reduce the amount of resources devoted to wine.

  Excise taxes that are intended to reduce the production and consumption of products with negative externalities are sometimes referred to as *sin taxes*. This name captures the idea that governments are motivated to impose these taxes to discourage activities that are perceived to be harmful or sinful. Excise taxes on cigarettes and alcohol in particular are commonly referred to as sin taxes.

QUICK REVIEW 18.3

- The benefits-received principle of taxation asserts that those who benefit from government services should pay the taxes needed to finance them; by contrast, the ability-to-pay principle asserts that taxes should be apportioned by income and wealth.
- A tax is (*a*) progressive if the average amount taxed away increases with income, (*b*) regressive if the average amount taxed away decreases with income, and (*c*) proportional if the average amount taxed away remains constant as income increases.
- Given fixed demand, more elastic supply shifts tax burdens to consumers; given fixed supply, more elastic demand shifts tax burdens to producers.
- Taxes imposed in markets raise the market equilibrium price, reduce the market equilibrium output, and normally generate efficiency losses.

Probable Incidence of U.S. Taxes

LO18.8 Discuss the probable incidence of U.S. taxes and how the distribution of income between rich and poor is affected by government taxes, transfers, and spending.

Let's look now at the probable incidence of each of the major sources of tax revenue in the United States.

Personal Income Tax

The incidence of the personal income tax generally is on the individual because there is little chance for shifting it. For every dollar paid to the tax, individuals have one less dollar in their pocketbooks. The same ordinarily holds true for inheritance taxes.

Payroll Taxes

As discussed earlier, employees and employers in 2012 *each* paid 7.65 percent in FICA taxes on a worker's annual earnings up to the 2012 Social Security cap of $110,100 and then 1.45 percent on any additional earnings.

Workers bear the full burden of their half of the Social Security and Medicare payroll taxes. As is true for the income tax, they cannot shift the payroll taxes that they pay to anyone else.

But what about the other half of the FICA tax that is levied on employers? Who pays that? The consensus view is that part of the employers' half of the FICA tax gets shifted to workers in the form of lower before-tax wages. By making it more costly to hire workers, the payroll tax reduces the demand for labor relative to supply. That reduces the market wages that employers pay workers. In a sense, employers "collect" some of the payroll tax they owe from their workers.

Corporate Income Tax

In the short run, the incidence of the corporate income tax falls on the company's stockholders (owners), who bear the burden of the tax through lower dividends or smaller amounts of retained corporate earnings. Here is why. A firm currently charging the profit-maximizing price and producing the profit-maximizing output will have no reason to change product price, output, or wages when a tax on corporate income (profit) is imposed. The price and output combination yielding the greatest profit before the tax will still yield the greatest profit after a fixed percentage of the firm's profit is removed by a corporate income tax. So, the company's stockholders will not be able to shift the tax to consumers or workers.

As previously indicated, the situation may be different in the long run. Workers, in general, may bear a significant part of the corporate income tax in the form of lower wage growth. Because it reduces the return on investment, the corporate income tax may slow the accumulation of capital (plant and equipment). It also may prompt some U.S. firms to relocate abroad in countries that have lower corporate tax rates. In either case, the tax may slow the growth of U.S. labor productivity, which depends on American workers having access to more and better equipment. We know from Figure 15.1 that the growth of labor productivity is the main reason labor demand grows over time. If the corporate income tax reduces the growth of labor productivity, then labor demand and wages may rise less rapidly. In this indirect way—and over long periods of time—workers may bear part of the corporate income tax.

Sales and Excise Taxes

A *sales tax* is a general excise tax levied on a full range of consumer goods and services, whereas a *specific excise tax* is one levied only on a particular product. Sales taxes are usually transparent to the buyer, whereas excise taxes are often "hidden" in the price of the product. But whether they are hidden or clearly visible, both are often partly or largely shifted to consumers through higher equilibrium product prices (as in Figures 18.7 through 18.9). Sales taxes and excise taxes may get shifted to different extents, however. Because a sales tax covers a much wider range of products than an excise tax, there is little chance for consumers to avoid the price boosts that sales taxes entail. They cannot reallocate their expenditures to untaxed, lower-priced products. Therefore, sales taxes tend to be shifted in their entirety from producers to consumers.

Excise taxes, however, fall on a select list of goods. Therefore, the possibility of consumers turning to substitute goods and services is greater. An excise tax on theater tickets that does not apply to other types of entertainment might be difficult to pass on to consumers via price increases. Why? The answer is provided in Figure 18.8a, where demand is elastic. A price boost to cover the excise tax on theater tickets might cause consumers to substitute alternative types of entertainment. The higher price would reduce sales so much that a seller would be better off to bear all, or a large portion of, the excise tax.

With other products, modest price increases to cover taxes may have smaller effects on sales. The excise taxes on gasoline, cigarettes, and alcoholic beverages provide examples. Here consumers have few good substitute products to which they can turn as prices rise. For these goods, the seller is better able to shift nearly all the excise tax to consumers. Example: Prices of cigarettes have gone up nearly in lockstep with the recent, substantial increases in excise taxes on cigarettes.

As indicated in Global Perspective 18.2, the United States depends less on sales and excise taxes for tax revenue than do several other nations.

Property Taxes

Many property taxes are borne by the property owner because there is no other party to whom they can be shifted. This is typically true for taxes on land, personal property, and owner-occupied residences. Even when land is sold, the property tax is not likely to be shifted. The buyer will understand that future taxes will have to be paid on it, and this expected taxation would be reflected in the price the buyer is willing to offer for the land.

Taxes on rented and business property are a different story. Taxes on rented property can be, and usually are, shifted wholly or in part from the owner to the tenant by the process of boosting the rent. Business property taxes are treated as a business cost and are taken into account in establishing product price; hence such taxes are ordinarily shifted to the firm's customers.

Table 18.2 (on the following page) summarizes this discussion of the shifting and incidence of taxes.

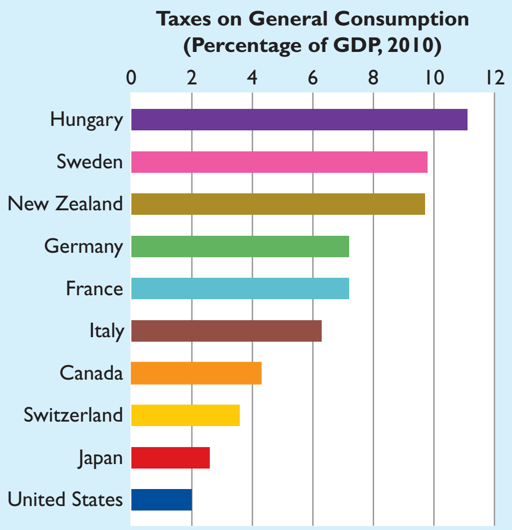
The U.S. Tax Structure

Is the overall U.S. tax structure—federal, state, and local taxes combined—progressive, proportional, or

GLOBAL PERSPECTIVE 18.2

Taxes on General Consumption as a Percentage of GDP, Selected Nations

A number of advanced industrial nations rely much more heavily on consumption taxes—sales taxes, specific excise taxes, and value-added taxes—than does the United States. A value-added tax, which the United States does not have, applies only to the difference between the value of a firm's sales and the value of its purchases from other firms. As a percentage of GDP, the highest tax rates on consumption are in countries that have value-added taxes.



Source: Organization for Economic Cooperation and Development, *OECD Stat Extracts*, **stats.oecd.org**.

regressive? The question is difficult to answer. Estimates of the distribution of the total tax burden depend on the extent to which the various taxes are shifted to others, and who bears the burden is subject to dispute. But the majority view of economists who study taxes is as follows:

- ***The federal tax system is progressive.*** Overall, higher-income groups pay larger percentages of their income as federal taxes than do lower-income groups. Although federal payroll taxes and excise taxes are regressive, the federal income tax is sufficiently progressive to make the overall federal tax system progressive. About one-third of federal income tax filers owe no tax at all. In fact, because of fully refundable tax credits designed to reduce poverty and promote work, millions of households receive tax rebates even though their income tax bill is zero. Most of the federal income tax is paid by higher-income taxpayers. In 2010 (the latest year for which data have been compiled), the top 1 percent of income-tax filers paid 37.4 percent of the federal income tax; the top 5 percent paid 59.1 percent of the tax.

  The overall progressivity of the federal tax system is confirmed by comparing effective (average) tax rates, which are found by dividing the total of federal income, payroll, and excise taxes paid at various income levels by the total incomes earned by the people at those various income levels. In 2009, the 20 percent of the households with the lowest income paid an effective tax rate of 1.0 percent. The 20 percent of households with the highest income paid a 23.2 percent rate. The top 1 percent paid a 28.9 percent rate.[1]
- ***The state and local tax structures are largely regressive.*** As a percentage of income, property taxes and sales taxes fall as income rises. Also, state income

[1] *The Distribution of Household Income and Federal Taxes, 2008 to 2009*, Congressional Budget Office, July 2012.

TABLE 18.2 The Probable Incidence of Taxes

Type of Tax	Probable Incidence
Personal income tax	The household or individual on which it is levied.
Payroll taxes	Workers pay the full tax levied on their earnings and part of the tax levied on their employers.
Corporate income tax	In the short run, the full tax falls on owners of the businesses. In the long run, some of the tax may be borne by workers through lower wages.
Sales tax	Consumers who buy the taxed products.
Specific excise taxes	Consumers, producers, or both, depending on elasticities of demand and supply.
Property taxes	Owners in the case of land and owner-occupied residences; tenants in the case of rented property; consumers in the case of business property.

LAST WORD

Taxation and Spending: Redistribution versus Recycling

Many Think of Taxes as the Best Way to Level the Income Distribution, but the Real Action Is in Expenditures.

Modern governments face substantial political pressure to ensure a fair distribution of society's economic benefits. In most people's minds, this boils down to taxing the rich more than the poor, which is why there is such a focus on whether particular taxes are progressive or regressive.

But taxing the rich cannot by itself alter the income distribution. One other thing is needed: The taxes taken from the rich have to flow to the poor. In particular, they have to be spent on goods, services, and programs that are used mostly by the poor rather than on goods, services, and programs that are used mostly by the rich. If government doesn't do this, the tax revenues of the rich will simply be recycled back to the rich rather than being redistributed to the poor.

Until recently, however, economists had only patchy evidence about whether our system of taxation and spending actually redistributes income from the rich to the poor. The problem was that the U.S. government only publishes statistics on whether the rich are being taxed more than the poor. It does not publish statistics on who receives most of its spending.

Fortunately, two economists from the nonpartisan Tax Foundation took it upon themselves to calculate those statistics. By combining data on government spending with household questionnaire responses in which people report what goods and services they consume, economists Andrew Chamberlain and Gerald Prante obtained the first reliable estimates of whether the government transfers significant amounts of income from the rich to the poor.*

As it turns out, the government *does* transfer an enormous amount of income from those with high incomes to those with low

incomes. Not only do people with high incomes pay a much larger fraction of their incomes in taxes, it is also the case that the majority of that money gets transferred to the poor because government spending is indeed concentrated on programs that are used more

*Andrew Chamberlain and Gerald Prante, "Who Pays Taxes and Who Receives Government Spending? An Analysis of Federal, State and Local Tax and Spending Distributions, 1991–2004," Tax Foundation Working Paper No. 1, 2007.

taxes are generally less progressive than the federal income tax.

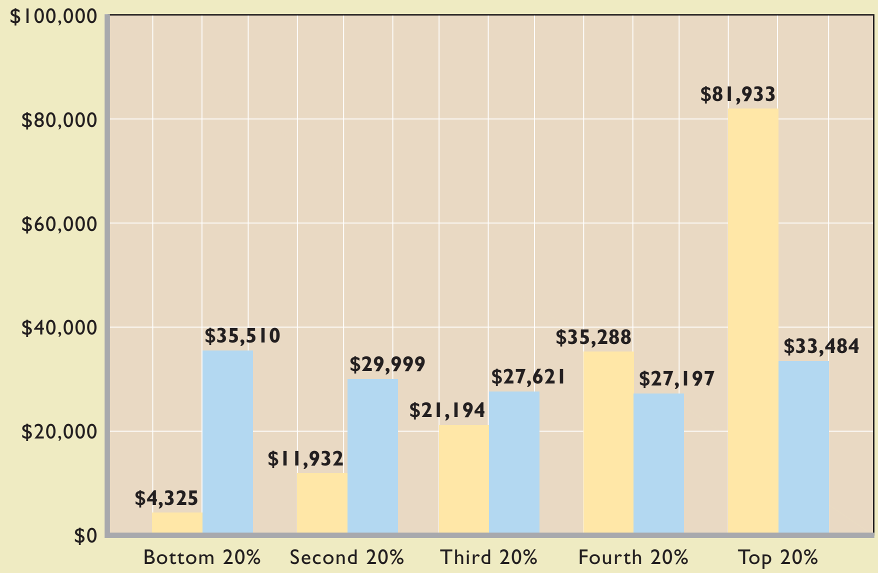
- ***The overall U.S. tax system is progressive.*** Higher-income people carry a substantially larger tax burden, as a percentage of their income, than do lower-income people, as discussed in this chapter's Last Word.
- ***The overall U.S. tax system is more progressive than that of other rich countries.*** A study by the Organization for Economic Cooperation and Development (OECD) concluded that the U.S. tax system is the most progressive among OECD nations and therefore more progressive than those of countries such as Canada, Japan, France, Sweden, Germany, Korea, Australia, and the United Kingdom.[2] This is because those other nations rely much more heavily on national sales taxes and value-added taxes, both of which are regressive.

This chapter's Last Word also points out that the income tax system cannot be relied upon by itself to substantially alter the distribution of income because the

[2]*Growing Unequal? Income Distribution and Poverty in OECD Countries*, Organization for Economic Cooperation and Development, 2008.

by the poor than by the rich. These include welfare, subsidized health care, public education, and jobs programs. The poor also benefit from government-provided public goods that are available to everyone on an equal basis—things like public roads, clean drinking water, national defense, and so on.

The size and impact of the income transfers from rich to poor are most clearly understood by looking at the nearby figure, which groups the 133 million households living in the United States in 2004 into one of five equally sized groups (or quintiles) on the basis of household income. The quintiles are labeled Bottom 20%, Second 20%, Third 20%, Fourth 20%, and Top 20%. The yellow and blue bars above each quintile show, respectively, how much in taxes its members paid on average and how much in government spending they received on average during 2004.



The first thing to notice is how much more money the poor received in government spending than they paid in taxes that year. A comparison of the yellow and blue bars for the bottom quintile reveals that the poorest households received $31,185 (= $35,510 in government spending − $4,325 in taxes) more in government spending than they paid in taxes in 2004. By contrast, households in the top 20 percent of the income distribution paid $48,449 more in taxes than they received in government spending that year.

This $48,449 per-household excess paid by households in the top quintile plus the $8,091 per-household excess paid by the households in the second-highest quintile provided the money that allowed the members of the lower three quintiles to receive more in government spending than they paid in taxes. In total, the transfers from the top two quintiles to the bottom three quintiles amounted to more than $1 trillion in 2004, or about 10 percent of all income earned by households that year.

The authors also found that the average tax rates paid by the five quintiles were, respectively, 13 percent for the bottom quintile, 23.2 percent for the second quintile, 28.2 percent for the third quintile, 31.3 percent for the fourth quintile, and 34.5 percent for the top quintile. Thus, the overall tax system is highly progressive (due to the federal income tax) despite many individual taxes being quite regressive.

But, more significantly, the spending made possible by taxing the rich more than the poor disproportionately flows back to the poor rather than being recycled to the rich. In fact, households in the top quintile receive back only 41 cents in government spending for each dollar they pay in taxes—which means that the remaining 59 cents are channeled to poorer households.

government might choose to spend the taxes collected from the rich to pay for things that are used more by the rich than the poor. In actual fact, however, this does not happen in the United States because the government uses a large portion of the tax revenues collected from the rich to make income transfer payments to the poor and to pay for the provision of goods and services that are utilized more by the poor than the rich. The transfer payments by themselves are so large that they almost quadruple the incomes of the poorest fifth of U.S. households. Thus, the combined tax-transfer system levels the income distribution by much more than the tax system does on its own.

QUICK REVIEW 18.4

- Some taxes are borne by those taxed while other taxes are shifted to someone else.
- The personal income tax and the corporate income tax (in the short run) are borne by those taxed.
- Sales taxes are shifted to consumers; the employer share of the payroll tax is shifted to workers; excise taxes may be shifted to consumers; and property taxes on rental properties are shifted to tenants.
- The federal tax structure is progressive. The state and local tax structures are regressive. The overall U.S. tax structure is progressive.

SUMMARY

LO18.1 Use a circular flow diagram to illustrate how the allocation of resources is affected by government's revenue and expenditure decisions.

The funds used to pay for government purchases and transfers come from taxes, proprietary income, and borrowing. The ability to borrow allows governments to maintain high spending during economic downturns, but government borrowing when the economy is doing well may "crowd out" private-sector investment.

LO18.2 Identify the main categories of government spending and the main sources of government revenue.

Government purchases exhaust (use up or absorb) resources; transfer payments do not. Government purchases have declined from about 22 percent of domestic output in 1960 to about 20 percent today. Transfer payments, however, have grown rapidly. As a percentage of GDP, total government spending (purchases plus transfers) now stands at about 35 percent, up from 27 percent in 1960.

LO18.3 List the main categories of federal revenue and spending and describe the difference between marginal and average tax rates.

The main categories of federal spending are pensions and income security, national defense, health, and interest on the public debt; federal revenues come primarily from personal income taxes, payroll taxes, and corporate income taxes.

It is important to distinguish between marginal and average tax rates. An income tax system may have different rates for different ranges of income. The tax rate that applies to any particular range is that range's marginal tax rate. The average tax rate is found by dividing the total amount of taxes paid on the full value of a taxpayer's taxable income and then dividing by the amount of the taxable income.

LO18.4 List the main categories of state and local revenue and spending.

States derive their revenue primarily from sales and excise taxes and personal income taxes; major state expenditures go to education, public welfare, health and hospitals, and highways. Local communities derive most of their revenue from property taxes; education is their most important expenditure. State and local tax revenues are supplemented by sizable revenue grants from the federal government.

LO18.5 Discuss the magnitude and distribution across job categories of government employment at the local, state, and federal levels.

Slightly over half of state and local government employees work in education. Just over half of federal government employees work for either the postal service or in national defense.

LO18.6 Summarize the different philosophies regarding the distribution of a nation's tax burden.

The benefits-received principle of taxation states that those who receive the benefits of goods and services provided by government should pay the taxes required to finance them. The ability-to-pay principle states that those who have greater income should be taxed more, absolutely and relatively, than those who have less income.

LO18.7 Explain the principles relating to tax shifting, tax incidence, and the efficiency losses causes by taxes.

Excise taxes affect supply and therefore equilibrium price and quantity. The more inelastic the demand for a product, the greater is the portion of an excise tax that is borne by consumers. The greater the inelasticity of supply, the larger is the portion of the tax that is borne by the seller.

Taxation involves the loss of some output whose marginal benefit exceeds its marginal cost. The more elastic the supply and demand curves, the greater is the efficiency loss (or deadweight loss) resulting from a particular tax.

LO18.8 Discuss the probable incidence of U.S. taxes and how the distribution of income between rich and poor is affected by government taxes, transfers, and spending.

The federal personal income tax is progressive. The corporate income tax is roughly proportional. General sales, excise, payroll, and property taxes are regressive.

Some taxes are borne by those taxed; other taxes are shifted to someone else. The income tax, the payroll tax levied on workers, and the corporate income tax (in the short run) are borne by those taxed. In contrast, sales taxes are shifted to consumers, part of the payroll tax levied on employers is shifted to workers, and, in the long run, part of the corporate income tax is shifted to workers. Specific excise taxes may or may not be shifted to consumers, depending on the elasticities of demand and supply. Property taxes on owner-occupied property are borne by the owner; those on rental property are borne by tenants.

The federal tax structure is progressive; the state and local tax structures are regressive; and the overall tax structure is progressive.

As discussed in the Last Word, the overall tax-spending system in the United States redistributes significant amounts of income from high-income individuals to low-income individuals. Because of the highly progressive federal income tax, the overall tax system is progressive. In addition, spending flows disproportionately to those with lower incomes, so that the tax collections from the rich are redistributed to the poor rather than being recycled back to the rich.

TERMS AND CONCEPTS

government purchases

transfer payments

personal income tax

marginal tax rate

average tax rate

payroll taxes

corporate income tax
sales and excise taxes
property taxes
benefits-received principle
ability-to-pay principle
progressive tax
regressive tax
proportional tax
tax incidence
efficiency loss of a tax

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Use a circular flow diagram to show how the allocation of resources and the distribution of income are affected by each of the following government actions. **LO18.1**
   a. The construction of a new high school.
   b. A 2-percentage-point reduction of the corporate income tax.
   c. An expansion of preschool programs for disadvantaged children.
   d. The levying of an excise tax on polluters.
2. What do economists mean when they say government purchases are "exhaustive" expenditures whereas government transfer payments are "nonexhaustive" expenditures? Cite an example of a government purchase and a government transfer payment. **LO18.1**
3. What are the main categories of government spending? What are the main categories of government revenue? **LO18.2**
4. What is the most important source of revenue and the major type of expenditure at the federal level? **LO18.3**
5. For state and local governments, what are the three most important sources of revenue and types of expenditure? **LO18.4**
6. How do the top two categories of federal employment differ from the top two categories of local and state employment? **LO18.5**
7. Distinguish between the benefits-received and the ability-to-pay principles of taxation. Which philosophy is more evident in our present tax structure? Justify your answer. To which principle of taxation do you subscribe? Why? **LO18.6**
8. What is meant by a progressive tax? A regressive tax? A proportional tax? Comment on the progressivity or regressivity of each of the following taxes, indicating in each case where you think the tax incidence lies: (*a*) the federal personal income tax, (*b*) a 4 percent state general sales tax, (*c*) a federal excise tax on automobile tires, (*d*) a municipal property tax on real estate, (*e*) the federal corporate income tax, (*f*) the portion of the payroll tax levied on employers. **LO18.6**
9. What is the tax incidence of an excise tax when demand is highly inelastic? Highly elastic? What effect does the elasticity of supply have on the incidence of an excise tax? What is the efficiency loss of a tax, and how does it relate to elasticity of demand and supply? **LO18.7**
10. Given the inelasticity of cigarette demand, discuss an excise tax on cigarettes in terms of efficiency loss and tax incidence. **LO18.7**
11. **ADVANCED ANALYSIS** Suppose the equation for the demand curve for some product X is $P = 8 - 0.6Q$ and the supply curve is $P = 2 + 0.4Q$. What are the equilibrium price and quantity? Now suppose an excise tax is imposed on X such that the new supply equation is $P = 4 + 0.4Q$. How much tax revenue will this excise tax yield the government? Graph the curves, and label the area of the graph that represents the tax collection "TC" and the area that represents the efficiency loss of the tax "EL." Briefly explain why area EL is the efficiency loss of the tax but TC is not. **LO18.7**
12. Is it possible for a country with a regressive tax system to have a tax-spending system that transfers resources from the rich to the poor? **LO18.8**
13. **LAST WORD** Does a progressive tax system by itself guarantee that resources will be redistributed from the rich to the poor? Explain. Is the *tax* system in the United States progressive, regressive, or proportional? Does the *tax-spending* system in the United States redistribute resources from higher-income earners to lower-income earners?

REVIEW QUESTIONS

1. The city of Joslyn has three sources of revenue: borrowing, proprietary income from running the local electric power utility, and taxes. If it received \$10 million from running the electric power utility and borrowed \$40 million, how much did it collect in taxes? **LO18.2**
   a. \$140 million.
   b. \$110 million.
   c. \$100 million.
   d. Nothing.
2. Suppose George made \$20,000 last year and that he lives in the country of Harmony. The way Harmony levies income taxes, each citizen must pay 10 percent in taxes on their first \$10,000 in earnings and then 50 percent in taxes on anything else they might earn. So given that George earned \$20,000 last year, his marginal tax rate on the last dollar he earns will be __________ and his average tax rate for his entire income will be _________________. **LO18.3**
   a. 50 percent; 50 percent.
   b. 50 percent; less than 50 percent.
   c. 10 percent; 50 percent.
   d. 10 percent; less than 50 percent.

3. The nation of Upstandia uses kroner for money and its tax code is such that a person making 100,000 kroner per year pays 40,000 kroner per year in income taxes; a person making 200,000 kroner per year pays 70,000 kroner per year in income taxes; and a person making 300,000 kroner per year pays 90,000 kroner per year in income taxes. Upstandia's income tax system is: **LO18.6**
   - a. Progressive.
   - b. Regressive.
   - c. Proportional.
4. Identify each of the following taxes as being either progressive or regressive. **LO18.6**
   - a. Personal income tax.
   - b. Sales taxes.
   - c. Payroll taxes.
   - d. Property taxes.
5. The efficiency loss of imposing an excise tax is due to: **LO18.7**
   - a. Paying a higher price per unit.
   - b. Producing and consuming fewer units.
6. True or false. The incidence of property taxes that are levied on rented houses and apartments is high—meaning that they are paid almost entirely by the landlords, who are billed by the government for those taxes. **LO18.8**

PROBLEMS

1. Suppose a tax is such that an individual with an income of $10,000 pays $2,000 of tax, a person with an income of $20,000 pays $3,000 of tax, a person with an income of $30,000 pays $4,000 of tax, and so forth. What is each person's average tax rate? Is this tax regressive, proportional, or progressive? **LO18.7**
2. Suppose in Fiscalville there is no tax on the first $10,000 of income, but a 20 percent tax on earnings between $10,000 and $20,000 and a 30 percent tax on income between $20,000 and $30,000. Any income above $30,000 is taxed at 40 percent. If your income is $50,000, how much will you pay in taxes? Determine your marginal and average tax rates. Is this a progressive tax? **LO18.7**
3. For tax purposes, "gross income" is all the money a person receives in a given year from any source. But income taxes are levied on "taxable income" rather than gross income. The difference between the two is the result of many exemptions and deductions. To see how they work, suppose you made $50,000 last year in wages, earned $10,000 from investments, and were given $5,000 as a gift by your grandmother. Also assume that you are a single parent with one small child living with you. **LO18.7**
   - a. What is your gross income?
   - b. Gifts of up to $13,000 per year from any person are not counted as taxable income. Also, the "personal exemption" allows you to reduce your taxable income by $3,650 for each member of your household. Given these exemptions, what is your taxable income?
   - c. Next, assume you paid $700 in interest on your student loans last year, put $2,000 into a health savings account (HSA), and deposited $4,000 into an individual retirement account (IRA). These expenditures are all *tax exempt*, meaning that any money spent on them reduces taxable income dollar-for-dollar. Knowing that fact, now what is your taxable income?
   - d. Next, you can either take the so-called standard deduction or apply for itemized deductions (which involve a lot of tedious paperwork). You opt for the standard deduction that allows you as head of your household to exempt another $8,500 from your taxable income. Taking that into account, what is your taxable income?
   - e. Apply the tax rates shown in Table 18.1 to your taxable income. How much federal income tax will you owe? What is the marginal tax rate that applies to your last dollar of taxable income?
   - f. As the parent of a dependent child, you qualify for the government's $1,000 per-child "tax credit." Like all tax credits, this $1,000 credit "pays" for $1,000 of whatever amount of tax you owe. Given this credit, how much money will you actually have to pay in taxes? Using that actual amount, what is your average tax rate relative to your taxable income? What about your average tax rate relative to your gross income?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **SIX**

MICROECONOMIC ISSUES AND POLICIES

CHAPTER **19**

Antitrust Policy and Regulation

Learning Objectives

LO19.1 **List and explain the core elements of the major antitrust (antimonopoly) laws in the United States.**

LO19.2 **Describe some of the key issues relating to the interpretation and application of antitrust laws.**

LO19.3 **Identify and explain the economic principles and difficulties relating to the setting of prices (rates) charged by so-called natural monopolies.**

LO19.4 **Discuss the nature of "social regulation," its benefits and costs, and its optimal level.**

We now can apply the economics of product markets (Part 4) and the economics of resource markets and governments (Part 5) to selected microeconomic issues and policies.

In this chapter we look at three sets of government policies toward business: antitrust policy, industrial regulation, and social regulation. **Antitrust policy** consists of laws and government actions designed to prevent monopoly and promote competition. **Industrial regulation** pertains to government regulation of firms' prices (or "rates") within selected industries. **Social regulation** is government regulation of the conditions under which goods are produced, the physical characteristics of goods, and the impact of the production and consumption of goods on society.

Then, in the remaining four chapters of Part 6, we discuss issues and policies relating to agriculture, income inequality, health care, and immigration.

The Antitrust Laws

LO19.1 List and explain the core elements of the major antitrust (antimonopoly) laws in the United States.

The underlying purpose of antitrust policy (antimonopoly policy) is to prevent monopolization, promote competition, and achieve allocative efficiency. Although all economists would agree that these are meritorious goals, there is sharp conflict of opinion about the appropriateness and effectiveness of U.S. antitrust policy. As we will see, antitrust policy over the years has been neither clear-cut nor consistent.

Historical Background

Just after the U.S. Civil War (1861–1865), local markets widened into national markets because of improved transportation facilities, mechanized production methods, and sophisticated corporate structures. In the 1870s and 1880s, dominant firms formed in several industries, including petroleum, meatpacking, railroads, sugar, lead, coal, whiskey, and tobacco. Some of these oligopolists, near-monopolists, or monopolists were known as *trusts*—business combinations that assign control to a single decision group ("trustees"). Because these trusts "monopolized" industries, the word "trust" became synonymous with "monopoly" in common usage. The public, government, and historians began to define a business monopoly as a large-scale dominant seller, even though that seller was not always "a sole seller" as specified in the model of pure monopoly.

These dominant firms often used questionable tactics in consolidating their industries and then charged high prices to customers and extracted price concessions from resource suppliers. Farmers and owners of small businesses were particularly vulnerable to the power of large corporate monopolies and were among the first to oppose them. Consumers, labor unions, and economists were not far behind in their opposition.

The main economic case against monopoly is familiar to you from Chapter 12. Monopolists tend to produce less output and charge higher prices than would be the case if their industries were competitive. With pure competition, each competitive firm maximizes profit by producing the output level at which $P = \text{MC}$. That output level generates allocative efficiency because price P measures the marginal benefit to society of an extra unit of output while marginal cost MC reflects the cost of an extra unit. When $P = \text{MC}$, society cannot gain by producing 1 more or 1 less unit of the product. In contrast, a monopolist maximizes profit by producing the lower output level at which marginal revenue (rather than price) equals marginal cost. At this $\text{MR} = \text{MC}$ point, price exceeds marginal cost, meaning that society would obtain more benefit than it would incur cost by producing extra units. An underallocation of resources to the monopolized product occurs, and the economy suffers an efficiency loss. So society's economic well-being is less than it would be with greater competition.

But an efficiency loss isn't the only consequence of the monopolist's higher-than-competitive price. The higher price also transfers income from consumers to the monopolist. This transfer causes significant resentment because it results purely from the monopolist's ability to restrict output and cannot be justified on the basis of increased production costs. Consumers consequently express their ire to elected officials to "do something about the situation."

Responding to that pressure, government officials concluded in the late 1800s and early 1900s that monopolized industries lacked enough of the beneficial market forces that in competitive industries help to protect consumers, achieve fair competition, and achieve allocative efficiency. So the government instituted two alternative means of control as substitutes for, or supplements to, market forces:

- ***Regulatory agencies*** In the few markets where the nature of the product or technology creates a *natural monopoly*, the government established public regulatory agencies to control economic behavior.
- ***Antitrust laws*** In most other markets, government control took the form of antitrust (antimonopoly) legislation designed to inhibit or prevent the growth of monopoly.

Four particular pieces of federal legislation, as refined and extended by various amendments, constitute the basic law relating to monopoly structure and conduct.

Sherman Act of 1890

The public resentment of trusts that emerged in the 1870s and 1880s culminated in the **Sherman Act** of 1890. This cornerstone of antitrust legislation is surprisingly brief and, at first glance, directly to the point. The core of the act resides in two provisions:

- ***Section 1*** "Every contract, combination in the form of a trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations is declared to be illegal."
- ***Section 2*** "Every person who shall monopolize, or attempt to monopolize, or combine or conspire with any person or persons, to monopolize any part of the trade or commerce among the several states, or with foreign nations, shall be deemed guilty of a felony" (as later amended from "misdemeanor").

The Sherman Act thus outlawed *restraints of trade* (for example, collusive price-fixing and dividing up markets) as well as *monopolization*. Today, the U.S. Department of Justice, the Federal Trade Commission, injured private parties, or state attorney generals can file antitrust suits against alleged violators of the act. The courts can issue injunctions to prohibit anticompetitive practices or, if necessary, break up monopolists into competing firms. Courts can also fine and imprison violators. Further, parties injured by illegal combinations and conspiracies can sue the perpetrators for *treble damages*—awards of three times the amount of the monetary injury done to them.

The Sherman Act seemed to provide a sound foundation for positive government action against business monopolies. However, early court interpretations limited the scope of the act and created ambiguities of law. It became clear that a more explicit statement of the government's antitrust sentiments was needed. The business community itself sought a clearer statement of what was legal and what was illegal.

Clayton Act of 1914

The **Clayton Act** of 1914 contained the desired elaboration of the Sherman Act. Four sections of the act, in particular, were designed to strengthen and make explicit the intent of the Sherman Act:

- Section 2 outlaws *price discrimination* when such discrimination is not justified on the basis of cost differences and when it reduces competition.
- Section 3 prohibits **tying contracts,** in which a producer requires that a buyer purchase another (or others) of its products as a condition for obtaining a desired product.
- Section 7 prohibits the acquisition of stocks of competing corporations when the outcome would be less competition.
- Section 8 prohibits the formation of **interlocking directorates**—situations where a director of one firm is also a board member of a competing firm—in large corporations where the effect would be reduced competition.

The Clayton Act simply sharpened and clarified the general provisions of the Sherman Act. It also sought to outlaw the techniques that firms might use to develop monopoly power and, in that sense, was a preventive measure. Section 2 of the Sherman Act, by contrast, was aimed more at breaking up existing monopolies.

Federal Trade Commission Act of 1914

The **Federal Trade Commission Act** created the five-member Federal Trade Commission (FTC), which has joint federal responsibility with the U.S. Justice Department for enforcing the antitrust laws. The act gave the FTC the power to investigate unfair competitive practices on its own initiative or at the request of injured firms. It can hold public hearings on such complaints and, if necessary, issue **cease-and-desist orders** in cases where it discovers "unfair methods of competition in commerce."

The **Wheeler-Lea Act** of 1938 amended the Federal Trade Commission Act to give the FTC the additional responsibility of policing "deceptive acts or practices in commerce." In so doing, the FTC tries to protect the public against false or misleading advertising and the misrepresentation of products. So the Federal Trade Commission Act, as modified by the Wheeler-Lea Act, (1) established the FTC as an independent antitrust agency and (2) made unfair and deceptive sales practices illegal.

The FTC is highly active in enforcing the deceptive advertising statutes. As one recent example, in 2007 the FTC fined four makers of over-the-counter diet pills a collective $25 million for claiming their products produced fast and permanent weight loss.

Celler-Kefauver Act of 1950

The **Celler-Kefauver Act** amended the Clayton Act, Section 7, which prohibits a firm from merging with a competing firm (and thereby lessening competition) by acquiring its stock. Firms could evade Section 7, however, by instead acquiring the physical assets (plant and equipment) of competing firms. The Celler-Kefauver Act closed that loophole by prohibiting one firm from obtaining the physical assets of another firm when the effect would be reduced competition. Section 7 of the Clayton Act now prohibits anticompetitive mergers no matter how they are undertaken.

Antitrust Policy: Issues and Impacts

LO19.2 Describe some of the key issues relating to the interpretation and application of antitrust laws.

The effectiveness of any law depends on how the courts interpret it and on the vigor of government enforcement. The courts have been inconsistent in interpreting the antitrust laws. At times, they have applied them vigorously, adhering closely to their spirit and objectives. At other times, their interpretations have rendered certain laws nearly powerless. The federal government itself has varied considerably in its aggressiveness in enforcing the antitrust laws. Some administrations have made tough antitrust enforcement a high priority. Other administrations have taken a more laissez-faire approach, initiating few antitrust actions or even scaling back the budgets of the enforcement agencies.

Issues of Interpretation

Differences in judicial interpretations have led to vastly different applications of the antitrust laws. Two questions, in particular, have arisen: (1) Should the focus of antitrust policy be on monopoly behavior or on monopoly structure? (2) How broadly should markets be defined in antitrust cases?

Monopoly Behavior versus Monopoly Structure

A comparison of three landmark Supreme Court decisions reveals two distinct interpretations of Section 2 of the Sherman Act as it relates to monopoly behavior and structure.

In the 1911 **Standard Oil case,** the Supreme Court found Standard Oil guilty of monopolizing the petroleum industry through a series of abusive and anticompetitive actions. The Court's remedy was to divide Standard Oil into several competing firms. But the Standard Oil case left open an important question: Is every monopoly in violation of Section 2 of the Sherman Act or just those created or maintained by anticompetitive actions?

In the 1920 **U.S. Steel case,** the courts established the so-called **rule of reason,** under which not every monopoly is illegal. Only monopolies that "unreasonably" restrain trade violate Section 2 of the Sherman Act and are subject to antitrust action. Size alone is not an offense. Under the rule of reason, U.S. Steel was innocent of "monopolizing" because it had not resorted to illegal acts against competitors in obtaining and then maintaining its monopoly power. Unlike Standard Oil, which was a so-called bad trust, U.S. Steel was a "good trust" and therefore not in violation of the law.

In the **Alcoa case** of 1945 the courts touched off a 20-year turnabout. The Supreme Court sent the case to the U.S. court of appeals in New York because four of the Supreme Court justices had been involved with litigation of the case before their appointments. Led by Judge Learned Hand, the court of appeals held that, even though a firm's behavior might be legal, the mere possession of monopoly power (Alcoa held 90 percent of the aluminum ingot market) violated the antitrust laws. So Alcoa was found guilty of violating the Sherman Act.

These two cases point to a controversy in antitrust policy. Should a firm be judged by its behavior (as in the U.S. Steel case) or by its structure or market share (as in the Alcoa case)?

- "Structuralists" assume that any firm with a very high market share will behave like a monopoly. As a result, they assert that any firm with a very high market share is a legitimate target for antitrust action. Structuralists argue that changes in the structure of an industry, say, by splitting the monopolist into several smaller firms, will improve behavior and performance.
- "Behavioralists" assert that the relationship among structure, behavior, and performance is tenuous and unclear. They feel a monopolized or highly concentrated industry may be technologically progressive and have a good record of providing products of increasing quality at reasonable prices. If a firm has served society well and has engaged in no anticompetitive practices, it should not be accused of antitrust violation just because it has an extraordinarily large market share. That share may be the product of superior technology, superior products, and economies of scale. "Why use antitrust laws to penalize efficient, technologically progressive, well-managed firms?" they ask.

Over the past several decades, the courts have returned to the rule of reason first established in the 1920 U.S. Steel case, and most contemporary economists and antitrust enforcers reject strict structuralism. For instance, in 1982 the government dropped its 13-year-long monopolization case against IBM on the grounds that IBM had not unreasonably restrained trade, despite having possessed extremely high market share in the market for mainframe computers. More recently, the government has made no attempt to break up Intel's near monopoly in the sale of microprocessors for personal computers. And in prosecuting the Microsoft case (the subject of this chapter's Last Word), the federal government made it clear that the behavior used by Microsoft to maintain and extend its monopoly, not the presence of its large market share, violated the Sherman Act. In essence, the government declared Microsoft to be "a bad monopoly" that could become a good monopoly if it stopped doing bad things.

Defining the Relevant Market Courts often decide whether or not market power exists by considering the share of the market held by the dominant firm. They have roughly adhered to a "90-60-30 rule" in defining monopoly: If a firm has a 90 percent market share, it is definitely a monopolist; if it has a 60 percent market share, it probably is a monopolist; if it has a 30 percent market share, it clearly is not a monopolist. The market share will depend on how the market is defined. If the market is defined broadly to include a wide range of somewhat similar products, the firm's market share will appear small. If the market is defined narrowly to exclude such products, the market share will seem large. The Supreme Court has the final say on how broadly to define relevant markets, but the Supreme Court has not always been consistent.

In the Alcoa case, the Court used a narrow definition of the relevant market: the aluminum ingot market. But in the **DuPont cellophane case** of 1956 the Court defined the market very broadly. The government contended that DuPont, along with a licensee, controlled 100 percent of the cellophane market. But the Court accepted DuPont's contention that the relevant market included all "flexible packaging materials"—waxed paper, aluminum foil, and so forth, in addition to cellophane. Despite DuPont's monopoly in the "cellophane market," it controlled only 20 percent of the market for "flexible wrapping materials." The Court ruled that this did not constitute a monopoly.

Issues of Enforcement

Some U.S. presidential administrations have enforced the antitrust laws more strictly than others. The degree of federal antitrust enforcement makes a difference in the overall degree of antitrust action in the economy. It is true that individual firms can sue other firms under the antitrust laws. For example, in 2005 AMD—a maker of microprocessors—filed an antitrust suit against Intel, claiming that Intel was a monopolist that used anticompetitive business practices to thwart the growth of AMD's market share. But major antitrust suits often last years and are highly expensive. Injured parties therefore often look to the federal government to initiate and litigate such cases. Once the federal government gains a conviction, the injured parties no longer need to prove guilt and can simply sue the violator to obtain treble damages. In many cases, lack of federal antitrust action therefore means diminished legal action by firms.

Why might one administration enforce the antitrust laws more or less strictly than another? The main reason is differences in political philosophies about the market economy and the wisdom of intervention by government. There are two contrasting general perspectives on antitrust policy.

The *active antitrust perspective* is that competition is insufficient in some circumstances to achieve allocative efficiency and ensure fairness to consumers and competing firms. Firms occasionally use illegal tactics against competitors to dominate markets. In other instances, competitors collude to fix prices or merge to enhance their monopoly power. Active, strict enforcement of the antitrust laws is needed to stop illegal business practices, prevent anticompetitive mergers, and remedy monopoly. This type of government intervention maintains the viability and vibrancy of the market system and thus allows society to reap its full benefits. In this view, the antitrust authorities need to act much like the officials in a football game. They must observe the players, spot infractions, and enforce the rules.

In contrast, the *laissez-faire perspective* holds that antitrust intervention is largely unnecessary, particularly as it relates to monopoly. Economists holding this position view competition as a long-run dynamic process in which firms battle against each other for dominance of markets. In some markets, a firm successfully monopolizes the market, usually because of its superior innovativeness or business skill. But in exploiting its monopoly power to raise prices, these firms inadvertently create profit incentives and profit opportunities for other entrepreneurs and firms to develop alternative technologies and new products to better serve consumers. As discussed in Chapter 2 and expanded upon in Chapter 11, a process of *creative destruction* occurs in which today's monopolies are eroded and eventually destroyed by tomorrow's technologies and products. The government therefore should not try to break up a monopoly. It should stand aside and allow the long-run competitive process to work.

The extent to which a particular administration adheres to—or leans toward—one of these contrasting antitrust perspectives usually gets reflected in the appointments to the agencies overseeing antitrust policy. Those appointees help determine how strictly the laws are enforced.

ORIGIN OF THE IDEA

O19.1

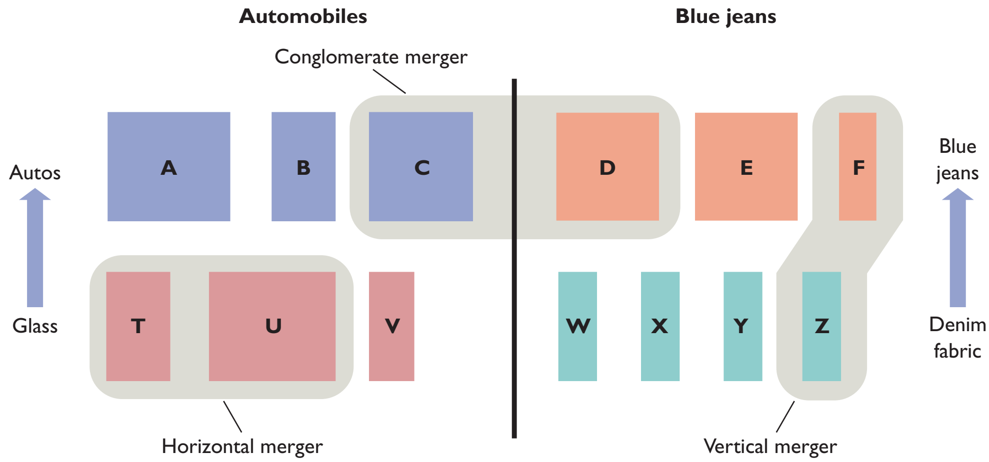
Creative destruction

Effectiveness of Antitrust Laws

Have the antitrust laws been effective? Although this question is difficult to answer, we can at least observe how the laws have been applied to monopoly, mergers, price-fixing, price discrimination, and tying contracts.

Monopoly On the basis of the rule of reason, the government has generally been lenient in applying antitrust laws to monopolies that have developed naturally. Generally, a firm will be sued by the federal government only if it has a very high market share and there is evidence of abusive conduct in achieving, maintaining, or extending its market dominance.

FIGURE 19.1 Types of mergers. Horizontal mergers (T + U) bring together firms selling the same product in the same geographic market; vertical mergers (F + Z) connect firms having a buyer-seller relationship; and conglomerate mergers (C + D) join firms in different industries or firms operating in different geographic areas.



But even if the federal government wins the antitrust lawsuit, there is still the matter of *remedy:* What actions should the court order to correct for the anticompetitive practices of the monopoly that lost the lawsuit?

The issue of remedy arose in two particularly noteworthy monopoly cases. The first was the AT&T (American Telephone and Telegraph) case in which the government charged AT&T with violating the Sherman Act by engaging in anticompetitive practices designed to maintain its domestic telephone monopoly. As part of an out-of-court settlement between the government and AT&T, in 1982 AT&T agreed to divest itself of its 22 regional telephone-operating companies.

A second significant monopoly case was the **Microsoft case.** In 2000 Microsoft was found guilty of violating the Sherman Act by taking several unlawful actions designed to maintain its monopoly of operating systems for personal computers. A lower court ordered that Microsoft be split into two competing firms. A court of appeals upheld the lower-court finding of abusive monopoly but rescinded the breakup of Microsoft. Instead of the structural remedy, the eventual outcome was a behavioral remedy in which Microsoft was prohibited from engaging in a set of specific anticompetitive business practices.

The antitrust agency for the European Union (EU) has generally been more aggressive than the United States in prosecuting monopolists. For example, in 2004 the EU fined Microsoft $600 million for monopolization and required it to share its computer code with other firms that supplied Windows applications (such as media players). The purpose of the code-sharing requirement was to enable competitors to compete on an equal footing against Microsoft's own application software. Microsoft was fined a further $1.3 billion by the EU in 2008 for failing to quickly and fully comply with the 2004 remedy ordering Microsoft to share its computer code.

Mergers The treatment of mergers, or combinations of existing firms, varies with the type of merger and its effect on competition.

Merger Types There are three basic types of mergers, as represented in Figure 19.1. This figure shows two stages of production (the input stage and the output, or final-product, stage) for two distinct final-goods industries (autos and blue jeans). Each rectangle (A, B, C, . . . X, Y, Z) represents a particular firm.

A **horizontal merger** is a merger between two competitors that sell similar products in the same geographic market. In Figure 19.1 this type of merger is shown as a combination of glass producers T and U. Actual examples of such mergers include Chase Manhattan's merger with Chemical Bank, Boeing's merger with McDonnell Douglas, and Exxon's merger with Mobil.

A **vertical merger** is a merger between firms at different stages of the production process. In Figure 19.1, the merger between firm Z, a producer of denim fabric, and firm F, a producer of blue jeans, is a vertical merger. Vertical mergers are mergers between firms that have buyer-seller relationships. Actual examples of such mergers are PepsiCo's mergers with Pizza Hut, Taco Bell, and Kentucky Fried Chicken. PepsiCo supplies soft drinks to each of these fast-food outlets. (In 1997, PepsiCo spun off these entities into a separate company now called Yum! Brands.)

A **conglomerate merger** is officially defined as any merger that is not horizontal or vertical; in general, it is the combination of firms in different industries or firms operating in different geographic areas. Conglomerate mergers can extend the line of products sold, extend the territory in which products are sold, or combine totally unrelated companies. In Figure 19.1, the merger between firm C, an auto manufacturer, and firm D, a blue jeans producer, is a conglomerate merger. Real-world examples of conglomerate mergers include the merger between Walt Disney Company (movies) and the American Broadcasting Company (radio and television) and the merger between America Online (Internet service provider) and Time Warner (communications).

Merger Guidelines: The Herfindahl Index The federal government has established very loose merger guidelines based on the Herfindahl index. Recall from Chapter 13 that this measure of concentration is the sum of the squared percentage market shares of the firms within an industry. An industry of only four firms, each with a 25 percent market share, has a Herfindahl index of 2,500 ($= 25^2 + 25^2 + 25^2 + 25^2$). In pure competition, where each firm's market share is minuscule, the index approaches 0 ($= 0^2 + 0^2 + \ldots + 0^2$). In pure monopoly, the index for that single firm is 10,000 ($= 100^2$).

The U.S. government uses Section 7 of the Clayton Act to block horizontal mergers that will substantially lessen competition. It is likely to challenge a horizontal merger if the postmerger Herfindahl index would be high (above 1,800) and if the merger has substantially increased the index (added 100 or more points). However, other factors, such as economies of scale, the degree of foreign competition, and the ease of entry of new firms, are also considered. Furthermore, horizontal mergers are usually allowed if one of the merging firms is suffering major and continuing losses. (This is one reason Boeing was allowed to acquire McDonnell Douglas in 1996: MD was losing money in producing its commercial airplanes.)

During the past several decades, the federal government has blocked several proposed horizontal mergers. For example, it blocked mergers between Staples and Office Depot, two major office-supply retailers; WorldCom and Sprint, two competing telecommunications firms; and Hughes (DirecTV) and Echostar (DISH Network), providers of direct-broadcast satellite television.

More recently, the federal government successfully challenged mergers between Snyder's of Hanover and Utz Quality Foods, makers of pretzels; Polypore and Microporous, battery-parts makers; and Blue Cross Blue Shield of Michigan and Physician's Health Plan of Michigan, health insurance providers.

Most *vertical mergers* escape antitrust prosecution because they do not substantially lessen competition in either of the two markets. (In Figure 19.1 neither the Herfindahl index in the industry producing denim fabric nor the index in the blue jeans industry changes when firms Z and F merge vertically.) However, a vertical merger between large firms in highly concentrated industries may be challenged. For example, in 1999 the threat of FTC action spurred Barnes & Noble to abandon its merger with Ingram Book Company, the nation's largest book wholesaler. The merger would have enabled Barnes & Noble to set the wholesale price of books charged to its direct retail competitors such as Borders and Amazon.com.

Conglomerate mergers are generally permitted. If an auto manufacturer acquires a blue jeans producer, no antitrust action is likely since neither firm has increased its own market share as a result. That means the Herfindahl index remains unchanged in each industry.

Price-Fixing

Price-fixing among competitors is treated strictly. Evidence of price-fixing, even by small firms, will bring antitrust action, as will other collusive activities such as scheming to rig bids on government contracts or dividing up sales in a market. In antitrust law, these activities are known as **per se violations;** they are "in and of themselves" illegal, and therefore are *not* subject to the rule of reason. To gain a conviction, the government or other party making the charge need show only that there was a conspiracy to fix prices, rig bids, or divide up markets, not that the conspiracy succeeded or caused serious damage to other parties.

Price-fixing investigations and court actions are common. (See the Consider This box to the right).

Price Discrimination

Price discrimination is a common business practice that rarely reduces competition and therefore is rarely challenged by government. The exception occurs when a firm engages in price discrimination as part of a strategy to block entry or drive out competitors.

Tying Contracts

The federal government strictly enforces the prohibition of tying contracts, particularly when practiced by dominant firms. For example, it stopped movie distributors from forcing theaters to buy the projection rights to a full package of films as a condition of showing a blockbuster movie. Also, it prevented Kodak—the dominant maker of photographic film—from requiring that consumers process their film only through Kodak.

Conclusions So what can we conclude about the overall effectiveness of antitrust laws? Antitrust policy has not been very effective in restricting the rise of or in breaking up monopolies or oligopolies resulting from legally undertaken internal expansions of firms. But most economists do not deem that to be a flaw. The antitrust laws have been used more effectively against predatory or abusive monopoly, but that effectiveness has been diminished by the slow legal process and consequently long time between the filing of charges and the implementation of remedies. In contrast, antitrust policy *has* been effective in prosecuting price-fixing and tying contracts.

Most economists conclude that, overall, U.S. antitrust policy has been moderately effective in achieving its goal of promoting competition and efficiency. Much of the success of antitrust policy arises from its deterrent effect on price-fixing and anticompetitive mergers. Some economists, however, think that enforcement of antitrust laws has been too weak. Others believe that parts of U.S. antitrust policy are anachronistic in an era of rapidly changing technology that continuously undermines existing monopoly power.

CONSIDER THIS . . .

Of Catfish and Art (and Other Things in Common)

Examples of price-fixing are numerous. Here are just a few:

- In 2001 a court found the auction houses Sotheby's and Christie's guilty of conspiring over a 6-year period to set the same commission rates for sellers at auctions.
- In the early 2000s Samsung (South Korea), Hynix Semiconductor (South Korea), Infineon (Germany), and Micron (United States) were found to have fixed the price of dynamic random access memory chips (DRAMs), which are used in personal computers, printers, cell phones, and other electronic devices. In 2004 and 2005, the companies paid fines totaling $645 million.
- In 2007 British Airlines and Korean Air agreed to pay fines of $300 million each for conspiring to fix fuel surcharges on passenger tickets and cargo.
- Between 2008 and 2010, five manufacturers of liquid crystal displays (LCDs)—LG, Sharp, Hitachi, Chi Mei Optoelectronics, and Chunghwa Picture Tubes—were fined a total of over $860 million by the U.S. Justice Department for fixing the prices of the displays they sold to computer maker Dell Inc.
- In 2009, three international cargo airlines—Cargolux of Luxembourg, Nippon Cargo of Japan, and Asiana Airlines of South Korea—were fined $214 million by the U.S. Justice Department for conspiring to fix international airline cargo rates.
- In 2012, nine Japanese, German, and Swedish auto parts makers were fined a total of $790 million for conspiring to fix the prices of automobile heater control panels. Eleven corporate officers received jail sentences ranging from one to two years.

QUICK REVIEW 19.1

- The Sherman Act of 1890 outlaws restraints of trade and monopolization; the Clayton Act of 1914 as amended by the Celler-Kefauver Act of 1950 outlaws price discrimination (when anticompetitive), tying contracts, anticompetitive mergers, and interlocking directorates.
- The Federal Trade Commission Act of 1914 as bolstered by the Wheeler-Lea Act of 1938 created the Federal Trade Commission (FTC) and gave it authority to investigate unfair methods of competition and deceptive acts or practices in commerce.
- Currently, the courts judge monopoly using a "rule of reason," first established in the U.S. Steel case of 1920. Under this rule, only monopolists that achieve or maintain their status abusively are in violation of the Sherman Act.
- "Structuralists" say highly concentrated industries will behave like monopolists; "behaviorists" hold that the relationship between industry structure and firm behavior is uncertain.
- The degree of strictness of enforcement of antitrust laws depends on the general antitrust philosophy of each U.S. administration and its appointees.
- Government treats existing monopoly relatively leniently, as long as it is not abusive; blocks most horizontal mergers between dominant, profitable firms in highly concentrated industries; and vigorously prosecutes price-fixing and tying contracts.

Industrial Regulation

LO19.3 Identify and explain the economic principles and difficulties relating to the setting of prices (rates) charged by so-called natural monopolies.

Antitrust policy assumes that society will benefit if a monopoly is prevented from evolving or if it is dissolved where it already exists. We now return to a special situation in which there is an economic reason for an industry to be organized monopolistically.

Natural Monopoly

A **natural monopoly** exists when economies of scale are so extensive that a single firm can supply the entire market at a lower average total cost than could a number of competing firms. Clear-cut circumstances of natural monopoly are relatively rare, but such conditions exist for many *public utilities*, such as local electricity, water, and natural gas providers. As discussed in Chapter 12, large-scale operations in some cases are necessary to obtain low unit costs and a low product price. Where natural monopoly occurs, competition is uneconomical. If the market were divided among many producers, economies of scale would not be achieved and unit costs and prices would be higher than necessary.

There are two possible alternatives for promoting better economic outcomes where natural monopoly exists. One is public ownership, and the other is public regulation.

Public ownership or some approximation of it has been established in a few instances. Examples: the Postal Service, the Tennessee Valley Authority, and Amtrak at the national level and mass transit, water supply systems, and garbage collection at the local level.

But *public regulation*, or what economists call *industrial regulation*, has been the preferred option in the United States. In this type of regulation, government commissions regulate the prices (or "rates") charged by natural monopolists. Table 19.1 lists the two major federal regulatory commissions and their jurisdictions. It also notes that all 50 states have commissions that regulate the intrastate activities and "utility rates" of local natural monopolies.

The economic objective of industrial regulation is embodied in the **public interest theory of regulation.** In that theory, industrial regulation is necessary to keep a natural monopoly from charging monopoly prices and thus harming consumers and society. The goal of such regulation is to garner for society at least part of the cost reductions associated with natural monopoly while avoiding the restrictions of output and high prices associated with unregulated monopoly. If competition is inappropriate or impractical, society should allow or even encourage a monopoly but regulate its prices. Regulation should then be structured so that ratepayers benefit from the economies of scale—the lower per-unit costs—that natural monopolists are able to achieve.

TABLE 19.1 **The Main Regulatory Commissions Providing Industrial Regulation**

Commission (Year Established)	Jurisdiction
Federal Energy Regulatory Commission (1930)*	Electricity, gas, gas pipelines, oil pipelines, water-power sites
Federal Communications Commission (1934)	Telephones, television, cable television, radio, telegraph, CB radios, ham operators
State public utility commissions (various years)	Electricity, gas, telephones

*Originally called the Federal Power Commission; renamed in 1977.

In practice, regulators seek to establish rates that will cover production costs and yield a "fair" return to the enterprise. The goal is to set price equal to average total cost so that the regulated firm receives a normal profit, as described in the "Regulated Monopoly" section of Chapter 12. In particular, you should carefully review Figure 12.9.

Problems with Industrial Regulation

There is considerable disagreement on the effectiveness of industrial regulation. Let's examine two criticisms.

Costs and Inefficiency An unregulated firm has a strong incentive to reduce its costs at each level of output because that will increase its profit. The regulatory commission, however, confines the regulated firm to a normal profit or a "fair return" on the value of its assets. If a regulated firm lowers its operating costs, the rising profit eventually will lead the regulatory commission to require that the firm lower its rates in order to return its profits to normal. The regulated firm therefore has little or no incentive to reduce its operating costs.

Worse yet, higher costs do not result in lower profit. Because the regulatory commission must allow the public utility a fair return, the regulated monopolist can simply pass through higher production costs to consumers by charging higher rates. A regulated firm may reason that it might as well have high salaries for its workers, opulent working conditions for management, and the like, since the "return" is the same in percentage terms whether costs are minimized or not. So, although a natural monopoly reduces costs through economies of scale, industrial regulation fosters considerable X-inefficiency (Figure 12.7). Due to the absence of competition, the potential cost savings from natural monopoly may never actually materialize.

Perpetuating Monopoly A second general problem with industrial regulation is that it sometimes perpetuates monopoly long after the conditions of natural monopoly have ended.

Technological change often creates the potential for competition in some or even all portions of the regulated industry. Examples: Trucks began competing with railroads; transmission of voice and data by microwave and satellites began competing with transmission over telephone wires; satellite television began competing with cable television; and cell phones began competing with landline phones.

But spurred by the firms they regulate, commissions often protect the regulated firms from new competition by either blocking entry or extending regulation to competitors. Industrial regulation therefore may perpetuate a monopoly that is no longer a natural monopoly and would otherwise erode. Ordinary monopoly, protected by government, may supplant natural monopoly. If so, the regulated prices may exceed those that would occur with competition. The beneficiaries of outdated regulation are the regulated firms and their employees. The losers are consumers and the potential entrants.

Example: Regulation of the railroads by the Interstate Commerce Commission (ICC) was justified in the late 1800s and early 1900s. But by the 1930s, with the emergence of a network of highways, the trucking industry had seriously undermined the monopoly power of the railroads. That is, for the transport of many goods over many routes, railroad service was no longer a natural monopoly. At that time it would have been desirable to dismantle the ICC and let railroads and truckers, along with barges and airlines, compete with one another. Instead, in the 1930s the ICC extended regulation of rates to interstate truckers. The ICC remained in place until its elimination in 1996. It was eliminated because the deregulation of railroads and trucking in the late 1970s and early 1980s had made its work irrelevant.

Second example: Until recently, both long-distance telephone companies such as AT&T as well as cable-television providers such as Time Warner were prohibited from offering local telephone services in competition with regulated local and regional telephone companies. But the very fact that AT&T, Time Warner, and other firms wanted to compete with regulated monopolies calls into question whether those local providers are in fact natural monopolies or government-protected monopolies.

Legal Cartel Theory

In Chapter 13, we noted that a *cartel* is formed when a group of previously competing firms makes a formal agreement to cease competing. They either control the price of a product by establishing the amounts of output that each cartel member will produce or they divide up the overall market for the product geographically so that each firm becomes a monopolist within its assigned region.

Privately organized cartels are illegal in the United States. But, in some cases, government regulations can have effects on prices and competition very similar to cartels. Thus, firms may seek to be regulated if they believe that regulation will reduce competition and raise prices in the same way a private cartel would. In essence, they see regulation as a way of creating a legal, publicly sanctioned cartel.

This possibility is known as the **legal cartel theory of regulation.** In place of having socially minded officials forcing regulation on natural monopolies to protect consumers, holders of this view see practical politicians "supplying" regulation to local, regional, and national firms that fear the impact of competition on their profits or even on their long-term survival. These firms desire regulation because it yields a legal monopoly that can virtually guarantee a profit. Specifically, the regulatory commission performs such functions as blocking entry (for example, in local telephone service). Or, where there are several firms, the commission divides up the market much like an illegal cartel (for example, prior to airline deregulation, the Civil Aeronautics Board assigned routes to specific airlines). The commission may also restrict potential competition by enlarging the "cartel" (for example, the ICC's addition of trucking to its regulatory domain).

While private cartels are illegal and unstable and often break down, the special attraction of a government-sponsored cartel under the guise of regulation is that it endures. The legal cartel theory of regulation suggests that regulation is a form of regulatory capture that results from the rent-seeking activities of private firms and the desire of politicians to be responsive in order to win reelection (Chapter 5).

Proponents of the legal cartel theory of regulation note that the Interstate Commerce Commission was welcomed by the railroads and that the trucking and airline industries both supported the extension of ICC regulation to their industries, arguing that unregulated competition was severe and destructive.

Occupational licensing is a labor market application of the legal cartel theory. Certain occupational groups—barbers, dentists, hairstylists, interior designers, dietitians, lawyers—demand stringent licensing on the grounds that it protects the public from charlatans and quacks. But skeptics say the real reason may be to limit entry into the occupational group so that practitioners can receive monopoly incomes.

Deregulation

Beginning in the 1970s, evidence of inefficiency in regulated industries and the contention that the government was regulating potentially competitive industries contributed to a wave of deregulation. Since then, Congress and many state legislatures have passed legislation that has deregulated in varying degrees the airline, trucking, banking, railroad, natural gas, television, and electricity industries. Deregulation has also occurred in the telecommunications industry, where antitrust authorities dismantled the regulated monopoly known as the Bell System (AT&T). Deregulation in the 1970s and 1980s was one of the most extensive experiments in economic policy to take place during the last 50 years.

The overwhelming consensus among economists is that deregulation has produced large net benefits for consumers and society. Most of the gains from deregulation have occurred in three industries: airlines, railroads, and trucking. Airfares (adjusted for inflation) declined by about one-third, and airline safety has continued to improve. Trucking and railroad freight rates (again, adjusted for inflation) dropped by about one-half.

Significant efficiency gains were also realized in long-distance telecommunications, and there have been slight efficiency gains in cable television, stock brokerage services, and the natural gas industry. Moreover, deregulation has unleashed a wave of technological advances that have resulted in such new and improved products and services as fax machines, cellular phones, fiber-optic cable, microwave communication systems, and the Internet.

The most recent and perhaps controversial industry to be deregulated is electricity. Deregulation is relatively advanced at the wholesale level, where firms can buy and sell electricity at market prices. They are also free to build generating facilities and sell electricity to local electricity providers at unregulated prices. In addition, several states have deregulated retail prices and encouraged households and businesses to choose among available electricity suppliers. This competition has generally lowered electricity rates for consumers and enhanced allocative efficiency.

But deregulation suffered a severe setback in California, where wholesale electricity prices, but not retail rates, were deregulated. Wholesale electricity prices surged in 2001 when California experienced electricity shortages. Because they could not pass on wholesale price increases to consumers, California electric utilities suffered large financial losses. California then filed lawsuits against several energy-trading companies that allegedly manipulated electricity supplies to boost the wholesale price of electricity during the California energy crisis. One multibillion-dollar energy trader—Enron—collapsed in 2002 when federal investigators uncovered a pattern of questionable and fraudulent business and accounting practices.

The California deregulation debacle and the Enron collapse have muddied the overall assessment of electricity deregulation in the United States. California's policy of deregulating only wholesale prices clearly did not work. So, even enthusiastic supporters of deregulation are now careful to reflect on the details of how, exactly, an industry should be dergulated.

QUICK REVIEW 19.2

- Natural monopoly occurs where economies of scale are so extensive that only a single firm can produce the product at minimum average total cost.
- The public interest theory of regulation says that government must regulate natural monopolies to prevent abuses arising from monopoly power. Regulated firms, however, have less incentive than competitive firms to reduce costs. That is, regulated firms tend to be X-inefficient.
- The legal cartel theory of regulation suggests that some firms seek government regulation to reduce price competition and ensure stable profits.
- Deregulation initiated by government in the past several decades has yielded large annual efficiency gains for society.

Social Regulation

LO19.4 Discuss the nature of "social regulation," its benefits and costs, and its optimal level.

The industrial regulation discussed in the preceding section has focused on the regulation of prices (or rates) in natural monopolies. But in the early 1960s a new type of regulation began to emerge. This *social regulation* is concerned with the conditions under which goods and services are produced, the impact of production on society, and the physical qualities of the goods themselves.

The federal government carries out most of the social regulation, although states also play a role. In Table 19.2 we list the main federal regulatory commissions engaged in social regulation.

Distinguishing Features

Social regulation differs from industrial regulation in several ways.

First, social regulation applies to far more firms than does industrial regulation. Social regulation is often applied "across the board" to all industries and directly affects more producers than does industrial regulation. For instance, while the industrial regulation of the Federal Energy Regulatory Commission (FERC) applies to a relatively small number of firms, the rules and regulations issued by

TABLE 19.2 **The Main Federal Regulatory Commissions Providing Social Regulation**

Commission (Year Established)	Jurisdiction
Food and Drug Administration (1906)	Safety and effectiveness of food, drugs, and cosmetics
Equal Employment Opportunity Commission (1964)	Hiring, promotion, and discharge of workers
Occupational Safety and Health Administration (1971)	Industrial health and safety
Environmental Protection Agency (1972)	Air, water, and noise pollution
Consumer Product Safety Commission (1972)	Safety of consumer products
Consumer Financial Protection Bureau (2011)	Fairness and transparency in lending and other financial services

the Occupational Safety and Health Administration (OSHA) apply to firms in all industries.

Second, social regulation intrudes into the day-to-day production process to a greater extent than industrial regulation. While industrial regulation focuses on rates, costs, and profits, social regulation often dictates the design of products, the conditions of employment, and the nature of the production process. As examples, the Consumer Product Safety Commission (CPSC) regulates the design of potentially unsafe products, and the Environmental Protection Agency (EPA) regulates the amount of pollution allowed during production.

Finally, social regulation has expanded rapidly during the same period in which industrial regulation has waned. Between 1970 and 1980, the U.S. government created 20 new social regulatory agencies. More recently, Congress has established new social regulations to be enforced by existing regulatory agencies. For example, the Equal Employment Opportunity Commission, which is responsible for enforcing laws against workplace discrimination on the basis of race, gender, age, or religion, has been given the added duty of enforcing the Americans with Disabilities Act of 1990. Under this social regulation, firms must provide reasonable accommodations for qualified workers and job applicants with disabilities. Also, sellers must provide reasonable access for customers with disabilities.

The names of the regulatory agencies in Table 19.2 suggest the reasons for their creation and growth: As much of our society had achieved a fairly affluent standard of living by the 1960s, attention shifted to improvement in the nonmaterial quality of life. The new focus called for safer products, less pollution, improved working conditions, and greater equality of economic opportunity.

The Optimal Level of Social Regulation

While economists agree on the need for social regulation, they disagree on whether or not the current level of such regulation is optimal. Recall that an activity should be expanded as long as its marginal benefit (MB) exceeds its marginal cost (MC). If the MB of social regulation exceeds its MC, then there is too little social regulation. But if MC exceeds MB, there is too much (review Figure 4.9). Unfortunately, the marginal costs and benefits of social regulation are not always easy to measure. So ideology about the proper size and role of government often drives the debate over social regulation as much as, or perhaps more than, economic cost-benefit analysis.

In Support of Social Regulation Defenders of social regulation say that it has achieved notable successes and, overall, has greatly enhanced society's well-being. They point out that the problems that social regulation confronts are serious and substantial. According to the National Safety Council, about 5,000 workers die annually in job-related accidents and 3.7 million workers suffer injuries that force them to miss a day or more of work. Air pollution continues to cloud major U.S. cities, imposing large costs in terms of reduced property values and increased health care expense. Numerous children and adults die each year because of poorly designed or manufactured products (for example, car tires) or tainted food (for example, *E. coli* in beef). Discrimination against some ethnic and racial minorities, persons with disabilities, and older workers reduces their earnings and imposes heavy costs on society.

Proponents of social regulation acknowledge that social regulation is costly. But they correctly point out that a high "price" for something does not necessarily mean that it should not be purchased. They say that the appropriate economic test should be not whether the costs of social regulation are high or low but, rather, whether the benefits of social regulation exceed the costs. After decades of neglect, they further assert, society cannot expect to cleanse the environment, enhance the safety of the workplace, and promote economic opportunity for all without incurring substantial costs. So statements about the huge costs of social regulation are irrelevant, say defenders, since the benefits are even greater. The public often underestimates those benefits since they are more difficult to measure than costs and often become apparent only after some time has passed (for example, the benefits of reducing global warming).

Proponents of social regulation point to its many specific benefits. Here are just a few examples: It is estimated that highway fatalities would be 40 percent greater annually

LAST WORD

United States v. Microsoft

The Microsoft Antitrust Case Is the Most Significant Monopoly Case since the Breakup of AT&T in the Early 1980s.

The Charges In May 1998 the U.S. Justice Department (under President Clinton), 19 individual states, and the District of Columbia (hereafter, "the government") filed antitrust charges against Microsoft under the Sherman Antitrust Act. The government charged that Microsoft had violated Section 2 of the act through a series of unlawful actions designed to maintain its "Windows" monopoly. It also charged that some of that conduct violated Section 1 of the Sherman Act.

Microsoft denied the charges, arguing it had achieved its success through product innovation and lawful business practices. Microsoft contended it should not be penalized for its superior foresight, business acumen, and technological prowess. It also pointed out that its monopoly was highly transitory because of rapid technological advance.

The District Court Findings In June 2000 the district court ruled that the relevant market was software used to operate Intel-compatible personal computers (PCs). Microsoft's 95 percent share of that market clearly gave it monopoly power. The court pointed out, however, that being a monopoly is not illegal. The violation of the Sherman Act occurred because Microsoft used anticompetitive means to maintain its monopoly power.

According to the court, Microsoft feared that the success of Netscape's Navigator, which allowed people to browse the Internet, might allow Netscape to expand its software to include a competitive PC operating system—software that would threaten the Windows monopoly. It also feared that Sun's Internet applications of its Java programming language might eventually threaten Microsoft's Windows monopoly.

To counter these and similar threats, Microsoft illegally signed contracts with PC makers that required them to feature

in the absence of auto safety features mandated through regulation. Compliance with child safety-seat and seat belt laws has significantly reduced the auto fatality rate for small children. The national air quality standards set by law have been reached in nearly all parts of the nation for sulfur dioxide, nitrogen dioxide, and lead. Moreover, recent studies clearly link cleaner air, other things equal, with increases in the values of homes. Affirmative action regulations have increased the labor demand for racial and ethnic minorities and females. The use of childproof lids has resulted in a 90 percent decline in child deaths caused by accidental swallowing of poisonous substances.

Some defenders of social regulation say there are many remaining areas in which greater regulation would generate net benefits to society. For instance, some call for greater regulation of the meat, poultry, and seafood industries to improve food safety. Others favor greater regulation of health care organizations and insurance companies to ensure "patients' rights" for consumers of health care services. Still others say that more regulation is needed to ensure that violent movies, CDs, and video games are not marketed to children.

Advocates of social regulation say that the benefits of such regulation are well worth the considerable costs. The costs are simply the price we must pay to create a hospitable, sustainable, and just society.

Criticisms of Social Regulation Critics of social regulation contend that, in many instances, it has been expanded to the point where the marginal costs exceed the marginal benefits. In this view, society would achieve net benefits by cutting back on irritating social regulation. Critics say that many social regulation laws are poorly written, making regulatory objectives and standards difficult to understand. As a result, regulators pursue goals well beyond the original intent of the legislation. Businesses

Internet Explorer on the PC desktop and penalized companies that promoted software products that competed with Microsoft products. Moreover, it gave friendly companies coding that linked Windows to software applications and withheld such coding from companies featuring Netscape. Finally, under license from Sun, Microsoft developed Windows-related Java software that made Sun's own software incompatible with Windows.

The District Court Remedy The district court ordered Microsoft to split into two competing companies, one initially selling the Windows operating system and the other initially selling Microsoft applications (such as Word, Hotmail, MSN, PowerPoint, and Internet Explorer). Both companies would be free to develop new products that compete with each other, and both could derive those products from the intellectual property embodied in the common products existing at the time of divestiture.

The Appeals Court Ruling In late 2000 Microsoft appealed the district court decision to a U.S. court of appeals. In 2001 the higher court affirmed that Microsoft illegally maintained its monopoly but tossed out the district court's decision to break up Microsoft. It agreed with Microsoft that the company was denied due process during the penalty phase of the trial and concluded that the district court judge had displayed an appearance of bias by holding extensive interviews with the press. The appeals court sent the remedial phase of the case to a new district court judge to determine appropriate remedies. The appeals court also raised issues relating to the wisdom of a structural remedy.

The Final Settlement At the urging of the new district court judge, the federal government (under then-President George W. Bush) and Microsoft negotiated a proposed settlement. With minor modification, the settlement became the final court order in 2002. The breakup was rescinded and replaced with a behavioral remedy. It (1) prevents Microsoft from retaliating against any firm that is developing, selling, or using software that competes with Microsoft Windows or Internet Explorer or is shipping a personal computer that includes both Windows and a non-Microsoft operating system; (2) requires Microsoft to establish uniform royalty and licensing terms for computer manufacturers wanting to include Windows on their PCs; (3) requires that manufacturers be allowed to remove Microsoft icons and replace them with other icons on the Windows desktop; and (4) calls for Microsoft to provide technical information to other companies so that they can develop programs that work as well with Windows as Microsoft's own products.

The Microsoft actions and conviction have indirectly resulted in billions of dollars of fines and payouts by Microsoft. The main examples are: To AOL Time Warner (Netscape), \$750 million; to the European Commission, \$1.9 billion; to Sun Microsystems, \$1.6 billion; to Novell, \$536 million; to Burst.com, \$60 million; to Gateway, \$150 million; to InterTrust, \$440 million; to RealNetworks, \$761 million; and to IBM, \$850 million.

Source: United States v. Microsoft (District Court Conclusions of Law), April 2000; *United States v. Microsoft* (Court of Appeals), June 2001; *U.S. v. Microsoft* (Final Judgment), November 2002; and Reuters and Associated Press News Services.

complain that regulators often press for additional increments of improvement, unmindful of costs.

Also, decisions must often be made and rules formed on the basis of inadequate information. Examples: CPSC officials may make decisions about certain ingredients in products on the basis of limited laboratory experiments that suggest that those ingredients might cause cancer. Or government agencies may establish costly pollution standards to attack the global-warming problem without knowing for certain whether pollution is the main cause of the problem. These efforts, say critics, lead to excessive regulation of business.

Moreover, critics argue that social regulations produce many unintended and costly side effects. For instance, the federal gas mileage standard for automobiles has been blamed for an estimated 2,000 to 3,900 traffic deaths a year because auto manufacturers have reduced the weight of vehicles to meet the higher miles-per-gallon standards. Other things equal, drivers of lighter cars have a higher fatality rate than drivers of heavier vehicles.

Finally, opponents of social regulation say that the regulatory agencies may attract overzealous workers who are hostile toward the market system and "believe" too fervently in regulation. For example, some staff members of government agencies may see large corporations as "bad guys" who regularly cause pollution, provide inadequate safety for workers, deceive their customers, and generally abuse their power in the community. Such biases can lead to seemingly never-ending calls for still more regulation, rather than objective assessments of the costs and benefits of added regulation.

Two Reminders

The debate over the proper amount of social regulation will surely continue. By helping determine costs and benefits, economic analysis can lead to more informed discussions and to better decisions. In this regard, economic analysis provides pertinent reminders for both ardent supporters and ardent opponents of social regulation.

There Is No Free Lunch Fervent supporters of social regulation need to remember that "there is no free lunch." Social regulation can produce higher prices, stifle innovation, and reduce competition.

Social regulation raises product prices in two ways. It does so directly because compliance costs normally get passed on to consumers, and it does so indirectly by reducing labor productivity. Resources invested in making workplaces accessible to disabled workers, for example, are not available for investment in new machinery designed to increase output per worker. Where the wage rate is fixed, a drop in labor productivity increases the marginal and average total costs of production. In effect, the supply curve for the product shifts leftward, causing the price of the product to rise.

Social regulation may have a negative impact on the rate of innovation. Technological advance may be stifled by, say, the fear that a new plant will not meet EPA guidelines or that a new medicine will require years of testing before being approved by the Food and Drug Administration (FDA).

Social regulation may weaken competition since it usually places a relatively greater burden on small firms than on large ones. The costs of complying with social regulation are, in effect, fixed costs. Because smaller firms produce less output over which to distribute those costs, their compliance costs per unit of output put them at a competitive disadvantage with their larger rivals. Social regulation is more likely to force smaller firms out of business, thus contributing to the increased concentration of industry.

Finally, social regulation may prompt some U.S. firms to move their operations to countries in which the rules are not as burdensome and therefore production costs are lower.

Less Government Is Not Always Better Than More
On the opposite side of the issue, opponents of social regulation need to remember that less government is not always better than more government. While the market system is a powerful engine for producing goods and services and generating income, it has certain flaws and can camouflage certain abuses. Through appropriate amounts of social regulation, government can clearly increase economic efficiency and thus society's well-being. Ironically, by "taking the rough edges off of capitalism," social regulation may be a strong pro-capitalism force. Properly conceived and executed, social regulation helps maintain political support for the market system. Such support could quickly wane should there be a steady drumbeat of reports of unsafe workplaces, unsafe products, discriminatory hiring, choking pollution, deceived loan customers, and the like. Social regulation helps the market system deliver not only goods and services but also a "good society."

QUICK REVIEW 19.3

- Social regulation is concerned with the conditions under which goods and services are produced, the effects of production on society, and the physical characteristics of the goods themselves.
- Defenders of social regulation point to the benefits arising from policies that keep dangerous products from the marketplace, reduce workplace injuries and deaths, contribute to clean air and water, and reduce employment discrimination.
- Critics of social regulation say uneconomical policy goals, inadequate information, unintended side effects, and overzealous personnel create excessive regulation, for which regulatory costs exceed regulatory benefits.

SUMMARY

LO19.1 List and explain the core elements of the major antitrust (antimonopoly) laws in the United States.

The cornerstones of antitrust policy are the Sherman Act of 1890 and the Clayton Act of 1914. The Sherman Act specifies that "every contract, combination . . . or conspiracy in the restraint of interstate trade . . . is . . . illegal" and that any person who monopolizes or attempts to monopolize interstate trade is guilty of a felony.

If a company is found guilty of violating the antimonopoly provisions of the Sherman Act, the government can either break up the monopoly into competing firms (a structural remedy) or prohibit it from engaging in specific anticompetitive business practices (a behavioral remedy).

The Clayton Act was designed to bolster and make more explicit the provisions of the Sherman Act. It declares that price discrimination, tying contracts, intercorporate stock acquisitions, and interlocking directorates are illegal when their effect is to reduce competition.

The Federal Trade Commission Act of 1914 created the Federal Trade Commission to investigate antitrust violations and to prevent the use of "unfair methods of competition." The FTC Act was amended by the Wheeler-Lea Act of 1938 to outlaw false and deceptive representation of products to consumers. Empowered by cease-and-desist orders, the FTC serves as a watchdog agency over unfair, deceptive, or false claims made by firms about their own products or the products of their competitors.

The Celler-Kefauver Act of 1950 amended the Clayton Act of 1914 to prohibit one firm from acquiring the assets of another firm when doing so would substantially reduce competition.

LO19.2 Describe some of the key issues relating to the interpretation and application of antitrust laws.

Some of the key issues in applying antitrust laws include (a) determining whether an industry should be judged by its structure or by its behavior; (b) defining the scope and size of the dominant firm's market; and (c) deciding how strictly to enforce the antitrust laws.

The courts treat price-fixing among competitors as a *per se violation*, meaning that the conduct is illegal independently of whether the conspiracy causes harm. In contrast, a *rule of reason* is used to assess monopoly. Only monopolies that unreasonably (abusively) achieve or maintain their status violate the law. Antitrust officials are more likely to challenge price-fixing, tying contracts, and horizontal mergers than to try to break up existing monopolies. Nevertheless, antitrust suits by the federal government led to the breakup of the AT&T monopoly in the early 1980s.

LO19.3 Identify and explain the economic principles and difficulties relating to the setting of prices (rates) charged by so-called natural monopolies.

The objective of industrial regulation is to protect the public from the market power of natural monopolies by regulating prices and quality of service.

Critics of industrial regulation contend that it can lead to inefficiency and rising costs and that in many instances it constitutes a legal cartel for the regulated firms. Legislation passed in the late 1970s and the 1980s has brought about varying degrees of deregulation in the airline, trucking, banking, railroad, and television broadcasting industries.

Studies indicate that deregulation of airlines, railroads, trucking, and telecommunications is producing sizable annual gains to society through lower prices, lower costs, and increased output. Less certain is the effect of the more recent deregulation of the electricity industry.

LO19.4 Discuss the nature of "social regulation," its benefits and costs, and its optimal level.

Social regulation is concerned with product safety, working conditions, and the effects of production on society. Whereas industrial regulation is on the wane, social regulation continues to expand. The optimal amount of social regulation occurs where MB = MC.

People who support social regulation point to its numerous specific successes and assert that it has greatly enhanced society's well-being. Critics of social regulation contend that businesses are excessively regulated to the point where marginal costs exceed marginal benefits. They also say that social regulation often produces unintended and costly side effects.

TERMS AND CONCEPTS

antitrust policy	cease-and-desist order	Microsoft case
industrial regulation	Wheeler-Lea Act	horizontal merger
social regulation	Celler-Kefauver Act	vertical merger
Sherman Act	Standard Oil case	conglomerate merger
Clayton Act	U.S. Steel case	per se violations
tying contracts	rule of reason	natural monopoly
interlocking directorates	Alcoa case	public interest theory of regulation
Federal Trade Commission Act	DuPont cellophane case	legal cartel theory of regulation

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Both antitrust policy and industrial regulation deal with monopoly. What distinguishes the two approaches? How does government decide to use one form of remedy rather than the other? **LO19.1, LO19.3**
2. Describe the major provisions of the Sherman and Clayton acts. What government entities are responsible for enforcing those laws? Are firms permitted to initiate antitrust suits on their own against other firms? **LO19.1**

3. Contrast the outcomes of the Standard Oil and U.S. Steel cases. What was the main antitrust issue in the DuPont cellophane case? In what major way do the Microsoft and Standard Oil cases differ? LO19.2
4. Why might one administration interpret and enforce the antitrust laws more strictly than another? How might a change of administrations affect a major monopoly case in progress? LO19.2
5. Suppose a proposed merger of firms would simultaneously lessen competition and reduce unit costs through economies of scale. Do you think such a merger should be allowed? LO19.2
6. In the 1980s, PepsiCo Inc., which then had 28 percent of the soft-drink market, proposed to acquire the Seven-Up Company. Shortly thereafter, the Coca-Cola Company, with 39 percent of the market, indicated it wanted to acquire the Dr Pepper Company. Seven-Up and Dr Pepper each controlled about 7 percent of the market. In your judgment, was the government's decision to block these mergers appropriate? LO19.2
7. Why might a firm charged with violating the Clayton Act, Section 7, try arguing that the products sold by the merged firms are in separate markets? Why might a firm charged with violating Section 2 of the Sherman Act try convincing the court that none of its behavior in achieving and maintaining its monopoly was illegal? LO19.2
8. "The social desirability of any particular firm should be judged not on the basis of its market share but on the basis of its conduct and performance." Make a counterargument, referring to the monopoly model in your statement. LO19.2
9. What types of industries, if any, should be subjected to industrial regulation? What specific problems does industrial regulation entail? LO19.3
10. In view of the problems involved in regulating natural monopolies, compare socially optimal (marginal-cost) pricing and fair-return pricing by referring again to Figure 12.9. Assuming that a government subsidy might be used to cover any loss resulting from marginal-cost pricing, which pricing policy would you favor? Why? What problems might such a subsidy entail? LO19.3
11. How does social regulation differ from industrial regulation? What types of benefits and costs are associated with social regulation? LO19.4
12. Use economic analysis to explain why the optimal amount of product safety may be less than the amount that would totally eliminate the risk of accidents and deaths. Use automobiles as an example. LO19.4
13. **LAST WORD** Under what law and on what basis did the federal district court find Microsoft guilty of violating the antitrust laws? What was the initial district court's remedy? How did Microsoft fare with its appeal to the court of appeals? Was the final remedy in the case a structural remedy or a behavioral remedy?

REVIEW QUESTIONS

1. True or false. Under the "rule of reason" that was established by the Supreme Court in the U.S. Steel case, a monopoly seller should be found guilty of violating antitrust laws even if it is charging low prices to consumers and acting the same way a competitive firm would act. LO19.2
2. How would you expect antitrust authorities to react to: LO19.2
   a. A proposed merger of Ford and General Motors.
   b. Evidence of secret meetings by contractors to rig bids for highway construction projects.
   c. A proposed merger of a large shoe manufacturer and a chain of retail shoe stores.
   d. A proposed merger of a small life-insurance company and a regional candy manufacturer.
   e. An automobile rental firm that charges higher rates for last-minute rentals than for rentals reserved weeks in advance.
3. When confronted with a natural monopoly that restricts output and charges monopoly prices, the two methods that governments have for promoting better outcomes are: LO19.3
   a. Public ownership and public regulation.
   b. Sole proprietorships and public goods.
   c. Antitrust law and horizontal mergers.
   d. Creative destruction and laissez-faire.
4. Which of the following is the correct name for the idea that certain firms prefer government regulation because regulation shields them from the pressures of competition and, in effect, guarantees them a regulated profit. LO19.3
   a. The public interest theory of regulation.
   b. The structuralists' theory of monopoly.
   c. The legal cartel theory of regulation.
   d. The public regulation theory of natural monopoly.
5. True or false. Economists believe that social regulation is an exception to the MB = MC rule because social regulation should in every case extend as far as possible in order to ensure safe products, less pollution, and improved working conditions. LO19.4

PROBLEMS

1. Suppose that there are only three types of fruit sold in the United States. Annual sales are 1 million tons of blueberries, 5 million tons of strawberries, and 10 million tons of bananas. Suppose that of those total amounts, the Sunny Valley Fruit Company sells 900,000 tons of blueberries, 900,000 tons of strawberries, and 7.9 million tons of bananas. LO19.2

a. What is Sunny Valley's market share if the relevant market is blueberries? If a court applies the "90-60-30 rule" when considering just the blueberry market, would it rule that Sunny Valley is a monopoly?

b. What is Sunny Valley's market share if the relevant market is all types of berries? Would the court rule Sunny Valley to be a monopolist in that market?

c. What if the relevant market is all types of fruit? What is Sunny Valley's market share, and would the court consider Sunny Valley to be a monopolist?

2. Carrot Computers and its competitors purchase touch screens for their handheld computers from several suppliers. The six makers of touch screens have market shares of, respectively, 19 percent, 18 percent, 14 percent, 16 percent, 20 percent, and 13 percent. **LO19.2**

a. What is the Herfindahl index for the touch screen manufacturing industry?

b. By how much would a proposed merger between the two smallest touch screen makers increase the Herfindahl index? Would the government be likely to challenge that proposed merger?

c. If Carrot Computers horizontally merges with its competitor Blueberry Handhelds, by how much would the Herfindahl index change for the touch screen industry?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **20**

Agriculture: Economics and Policy

Learning Objectives

LO20.1 Explain why agricultural prices and farm income are unstable.

LO20.2 Discuss why there has been a huge employment exodus from agriculture to other U.S. industries over the past several decades.

LO20.3 Relate the rationale for farm subsidies and the economics and politics of price supports (price floors).

LO20.4 Describe major criticisms of the price-support system in agriculture.

LO20.5 List the main elements of existing federal farm policy.

If you eat, you are part of agriculture! In the United States, agriculture is important for a number of reasons. It is one of the nation's largest industries and major segments of it provide real-world examples of the pure-competition model developed in Chapters 10 and 11. Also, agriculture clearly shows the effects of government policies that interfere with supply and demand. Further, the industry provides excellent illustrations of Chapter 5's special-interest effect and rent-seeking behavior. Finally, it demonstrates the globalization of markets for farm commodities.

This chapter examines the circumstances in agriculture that have resulted in government intervention, the types and outcomes of government intervention, and recent major changes in farm policy.

Economics of Agriculture

LO20.1 Explain why agricultural prices and farm income are unstable.

Although economists refer to *the* agriculture industry, this segment of the economy is extremely diverse. Agriculture encompasses cattle ranches, fruit orchards, dairies, poultry plants, pig farms, grain farms, feed lots, vegetable plots, sugar-cane plantations, and much more. Some farm commodities (for example, soybeans and corn) are produced by thousands of individual farmers. Other farm commodities (such as poultry) are produced by just a handful of large firms. Some farm products (for example, wheat, milk, and sugar) are heavily subsidized through federal government programs; other farm products (such as fruits, nuts, and potatoes) have much less government support.

Moreover, agriculture includes both farm products, or **farm commodities** (for example, wheat, soybeans, cattle, and rice), and also **food products** (items sold through restaurants or grocery stores). Generally, the number of competing firms in the market diminishes as farm products are refined into commercial food products. Although thousands of ranches and farms raise cattle, only four firms (Tyson, Cargill, JBS, and National) account for about 80 percent of red meat produced at cattle slaughtering/meat packing plants. And thousands of farms grow tomatoes, but only three companies (Heinz, Del-Monte, and Hunt) make the bulk of the ketchup sold in the United States.

Our focus in this chapter will be on farm commodities (or farm products) and the farms and ranches that produce them. Farm commodities usually are sold in highly competitive markets, whereas food products tend to be sold in markets characterized by monopolistic competition or oligopoly.

Partly because of large government subsidies, farming remains a generally profitable industry. U.S. consumers allocate 13 percent of their spending to food, and farmers and ranchers receive about $291 billion of revenue annually from sales of crops and livestock. Over the years, however, American farmers have experienced severely fluctuating prices and periodically low incomes. Further, they have had to adjust to the reality that agriculture is a declining industry. The farm share of GDP has declined from about 7 percent in 1950 to 1 percent today.

Let's take a close look at both the short-run and long-run economics of U.S. agriculture.

The Short Run: Price and Income Instability

Price and income instability in agriculture results from (1) an inelastic demand for agricultural products, combined with (2) fluctuations in farm output and (3) shifts of the demand curve for farm products.

Inelastic Demand for Agricultural Products In industrially advanced economies, the price elasticity of demand for agricultural products is low. For agricultural products in the aggregate, the elasticity coefficient is between 0.20 and 0.25. These figures suggest that the prices of farm products would have to fall by 40 to 50 percent for consumers to increase their purchases by a mere 10 percent. Consumers apparently put a low value on additional farm output compared with the value they put on additional units of alternative goods.

Why is this so? Recall that the basic determinant of elasticity of demand is substitutability. When the price of one product falls, the consumer tends to substitute that product for other products whose prices have not fallen. But in relatively wealthy societies this "substitution effect" is very modest for food. Although people may eat more, they do not switch from three meals a day to, say, five or six meals a day in response to a decline in the relative prices of farm products. Real biological factors constrain an individual's capacity to substitute food for other products.

The inelasticity of agricultural demand is also related to diminishing marginal utility. In a high-income economy, the population is generally well fed and well clothed; it is relatively saturated with the food and fiber of agriculture. Additional farm products therefore are subject to rapidly diminishing marginal utility. So very large price cuts are needed to induce small increases in food and fiber consumption.

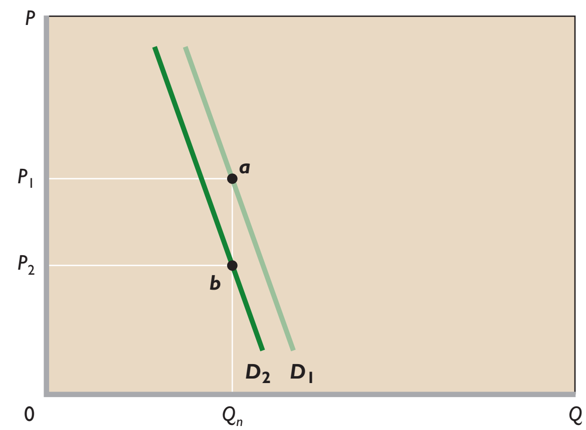
Fluctuations in Output Farm output tends to fluctuate from year to year, mainly because farmers have limited control over their output. Floods, droughts, unexpected frost, insect damage, and similar disasters can mean poor crops, while an excellent growing season means bumper crops (unusually large outputs). Such natural occurrences are beyond the control of farmers, yet they exert an important influence on output.

In addition to natural phenomena, the highly competitive nature of many parts of farming and ranching makes it difficult for those producers to form huge combinations to control production. If the thousands of widely scattered and independent producers happened to plant an unusually large or abnormally small portion of their land one year, an extra-large or a very small farm output would result even if the growing season were normal.

Curve *D* in Figure 20.1 illustrates the inelastic demand for agricultural products. Combining that inelastic demand with the instability of farm production, we can see why agricultural prices and incomes are unstable. Even if the

FIGURE 20.1 The effects of changes in farm output on agricultural prices and income. Because of the inelasticity of demand for farm products, a relatively small change in farm output (from Q_n to Q_p or Q_b) will cause a relatively large change in agricultural prices (from P_n to P_p or P_b). Farm income will change from the yellow area to the larger $0P_ppQ_p$ area or to the smaller $0P_bbQ_b$ area.

FIGURE 20.2 The effect of a demand shift on agricultural prices and income. Because of the highly inelastic demand for farm products, a small shift in demand (from D_1 to D_2) for farm products can drastically alter agricultural prices (P_1 to P_2) and farm income (area $0P_1aQ_n$ to area $0P_2bQ_n$), given a fixed level of production Q_n.



market demand for farm products remains fixed at D, its price inelasticity will magnify small changes in output into relatively large changes in agricultural prices and income. For example, suppose that a "normal" crop of Q_n results in a "normal" price of P_n and a "normal" farm income represented by the yellow rectangle. A bumper crop or a poor crop will cause large deviations from these normal prices and incomes because of the inelasticity of demand.

If a good growing season occurs, the resulting large crop of Q_b will reduce farm income to that of area $0P_bbQ_b$. When demand is inelastic, an increase in the quantity sold will be accompanied by a more-than-proportionate decline in price. The net result is that total revenue, that is, total farm income, will decline disproportionately.

Similarly, a small crop caused by, say, drought will boost total farm income to that represented by area $0P_ppQ_p$. A decline in output will cause more-than-proportionate increases in price and income when demand is inelastic. Ironically, for farmers as a group, a poor crop may be a blessing and a bumper crop a hardship.

Conclusion: With a stable market demand for farm products, the inelasticity of that demand will turn relatively small changes in output into relatively larger changes in agricultural prices and income.

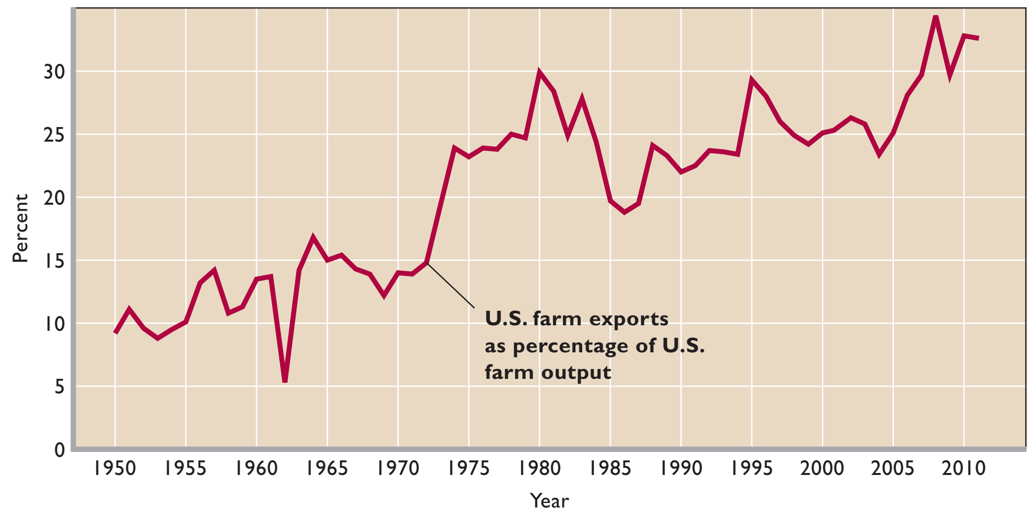
Fluctuations in Demand The third factor in the short-run instability of farm income results from shifts in the demand curve for agricultural products. Suppose that somehow farm output is stabilized at the "normal" level of Q_n in Figure 20.2. Now, because of the inelasticity of the demand for farm products, short-run changes in the demand for those products will cause markedly different prices and incomes to be associated with this fixed level of output.

A slight decline in demand from D_1 to D_2 will reduce farm income from area $0P_1aQ_n$ to $0P_2bQ_n$. So a relatively small decline in demand gives farmers significantly less income for the same amount of farm output. Conversely, a slight increase in demand—as from D_2 to D_1—provides a sizable increase in farm income for the same volume of output. Again, large price and income changes occur because demand is inelastic.

It is tempting to argue that the sharp declines in agricultural prices that accompany a decrease in demand will cause many farmers to close down in the short run, reducing total output and alleviating the price and income declines. But farm production is relatively insensitive to price changes in the short run because farmers' fixed costs are high compared with their variable costs.

Interest, rent, tax, and mortgage payments on land, buildings, and equipment are the major costs faced by the farmer. These are all fixed charges. Furthermore, the labor supply of farmers and their families can also be regarded as a fixed cost. As long as they stay on their farms, farmers cannot reduce their costs by firing themselves. Their variable costs are the costs of the small amounts of extra help they may employ, as well as expenditures for seed, fertilizer, and fuel. As a result of their high proportion of fixed costs, farmers are usually better off working their land even when they are losing money since they would lose much more by shutting down their operations for the year. Only in the long run will exiting the industry make sense for them.

FIGURE 20.3 U.S. farm exports as a percentage of farm output, 1950–2011. Exports of farm output have increased as a percentage of total farm output (the value of agricultural-sector production) in the United States. But this percentage has been quite variable, contributing to the instability of the demand for U.S. farm output.



Source: Derived by the authors from *Foreign Agricultural Trade of the United States,* **www.ers.usda.gov/Data/FATUS**; and Bureau of Economic Analysis, **www.bea.gov**.

But why is agricultural demand unstable? The major source of demand volatility in U.S. agriculture springs from its dependence on world markets. As we show in Figure 20.3, that dependency has increased since 1950. The yearly ups and downs of the line in the figure also reveal that, as a percentage of total U.S. farm output, farm exports are highly unstable.

The incomes of U.S. farmers are sensitive to changes in weather and crop production in other countries: Better crops abroad mean less foreign demand for U.S. farm products. Similarly, cyclical fluctuations in incomes in Europe or Southeast Asia, for example, may shift the demand for U.S. farm products. Changes in foreign economic policies may also change demand. For instance, if the nations of western Europe decide to provide their farmers with greater protection from foreign competition, U.S. farmers will have less access to those markets and demand for U.S. farm exports will fall.

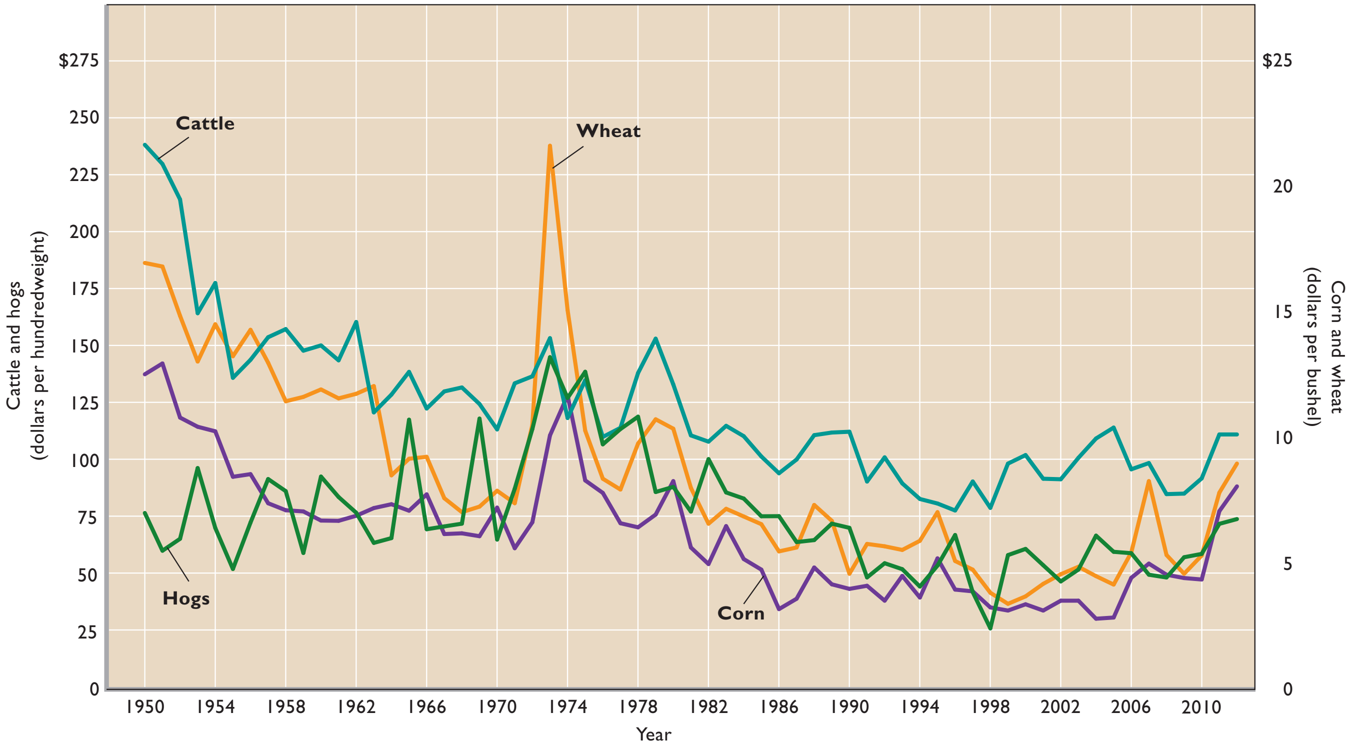
International politics also add to demand instability. Changing political relations between the United States and China and the United States and Russia have boosted exports to those countries in some periods and reduced them in others. Changes in the international value of the dollar may also be critical. Depreciation of the dollar increases the demand for U.S. farm products (which become cheaper to foreigners), whereas appreciation of the dollar diminishes foreign demand for U.S. farm products.

Figure 20.4 shows inflation-adjusted U.S. prices for cattle, hogs, corn, and wheat from 1950 through 2012. The short-run economics of price volatility is evident. So, too, is the general decline of real (inflation-adjusted) agricultural prices from 1950 through the late 1990s.

Since the late 1990s, however, the prices of these four commodities have not continued the overall downward trend that they had previously followed. Prices continued to be volatile but remained in 2009 about where they had been in the late 1990s. Then, major price spikes took place between 2009 and 2012. For example, the inflation-adjusted price of wheat more than doubled, increasing from $3.75 per bushel in 2009 to $8.31 per bushel in 2012.

The lack of a downward trend in commodity prices starting in the late 1990s coupled with the major price spikes that occurred between 2009 and 2012 led some economists to wonder whether the era of inflation-adjusted price decreases in agriculture had come to an end. They speculated that the rising demand for food in emerging economies such as China together with the growing demand for farm products to produce ethanol might reverse the long-run downward trend. But, as Figure 20.4 reveals, several previous price spikes as well as flat periods in agricultural prices have occurred. Each of these events eventually yielded to the general historical downward trend in agricultural prices. Whether this pattern will occur again remains to be seen.

FIGURE 20.4 Inflation-adjusted U.S. agricultural prices, selected commodities, 1950–2012. Inflation-adjusted U.S. prices (in 2005 dollars) for cattle, hogs, corn, and wheat since 1950 indicate both volatility and general decline.



Source: Author calculations using nominal values from Global Financial Data, **globalfinancialdata.com**, adjusted for inflation with the GDP deflator published by the Bureau of Economic Analysis, **bea.gov**.

The Long Run: A Declining Industry

LO20.2 Discuss why there has been a huge employment exodus from agriculture to other U.S. industries over the past several decades.

Two dynamic characteristics of agricultural markets explain why agriculture has been a declining industry:

- Over time, the supply of farm products has increased rapidly because of technological progress.
- The demand for farm products has increased slowly because it is inelastic with respect to income and because it is largely limited by population growth, which has not been rapid in the United States.

Let's examine each of these supply and demand forces.

Technology and Supply Increases

A rapid rate of technological advance has significantly increased the supply of agricultural products. This technological progress has many roots: the mechanization of farms, improved techniques of land management, soil conservation, irrigation, development of hybrid crops, availability of improved fertilizers and insecticides, polymer-coated seeds, and improvements in the breeding and care of livestock. The amount of capital used per farmworker increased 15 times between 1930 and 1980, permitting a fivefold increase in the amount of land cultivated per farmer. The simplest measure of these advances is the U.S. Agriculture Department's index of farm output per unit of farm labor. In 1950 a single unit of farm labor could produce 10 units of farm output. This amount increased to 30 in 1970, 42 in 1980, 64 in 1990, 90 in 2000, and 119 in 2009. Over the last half-century, productivity in agriculture has advanced twice as fast as productivity in the nonfarm economy.

Most of the technological advances in agriculture were not initiated by farmers. Rather, they are the result of government-sponsored programs of research and education and the initiative of the suppliers of farm inputs. Land-grant colleges, experiment stations, county agents of the Agricultural Extension Service, educational pamphlets issued by the U.S. Department of Agriculture (USDA), and the research departments of farm machinery, pesticide, and fertilizer producers have been the primary sources of technological advance in U.S. agriculture.

More recently, technological advance has been fueled by the incorporation of advanced information technologies

CONSIDER THIS . . .

Risky Business

The short-run instability of agricultural prices and farm income creates considerable risk in agriculture. Later in this chapter we will find that farm programs (direct payments, countercyclical payments, and "repay-or-default" loans) reduce the risk of farming for many farmers. But these programs are limited to certain crops, such as grains and oilseeds.

Fortunately, several private techniques for managing risk have become commonplace in agriculture. The purpose of these measures is to "smooth" income over time, "hedging" against short-run output and price fluctuations. Hedging is an action by a buyer or seller to protect against a change in future prices prior to an anticipated purchase or sale.

Farm risk-management techniques include:

- ***Futures markets.*** In the futures market, farmers can buy or sell farm products at prices fixed now, for delivery at a specified date in the future. If the price falls, farmers will still obtain revenue based on the higher price fixed in the futures market. If the price rises, the buyer will benefit by getting the farm commodity at the lower price fixed in the futures market.
- ***Contracting with processors.*** In advance of planting, farmers can directly contract with food processors (firms such as sugar beet refiners, ethanol plants, and feed lots) to assure themselves of a fixed price per unit of their farm or ranch output.
- ***Crop revenue insurance.*** Farmers can buy crop revenue insurance, which insures them against gross revenue losses resulting from storm damage and other natural occurrences.
- ***Leasing land.*** Farm operators can reduce their risk by leasing some of their land to other operators who pay them cash rent. The rent payment is stable, regardless of the quality of the crop and crop prices.
- ***Nonfarm income.*** Many farm households derive substantial parts of their total income from off-farm income, such as spousal work and agricultural investments. These more-stable elements of income cushion the instability of farm income.

Although farming remains a risky business, farm operators have found creative ways to manage the inherent risks of price and income instability.

into farming. Computers and the Internet give farmers instant access to information about soil conditions, estimated crop yields, farm-product prices, available land for purchase or lease, and much more. They also provide farmers with sophisticated business software to help track and manage their operations.

Lagging Demand

Increases in the demand for agricultural products have failed to keep pace with these technologically created increases in the supply of the products. The reason lies in the two major determinants of agricultural demand: income and population.

In developing countries, consumers must devote most of their meager incomes to agricultural products—food and clothing—to sustain themselves. But as income expands beyond subsistence and the problem of hunger diminishes, consumers increase their outlays on food at ever-declining rates. Once consumers' stomachs are filled, they turn to the amenities of life that manufacturing and services, not agriculture, provide. Economic growth in the United States has boosted average per capita income far beyond the level of subsistence. As a result, increases in the incomes of U.S. consumers now produce less-than-proportionate increases in spending on farm products.

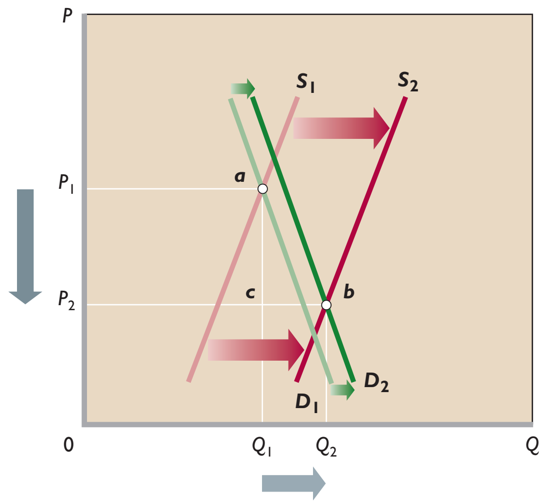
The demand for farm products in the United States is income-inelastic; it is quite insensitive to increases in income. Estimates indicate that a 10 percent increase in real per capita after-tax income produces about a 1 percent increase in consumption of farm products. That means a coefficient of income elasticity of 0.1 (= 0.01/0.10). So as the incomes of Americans rise, the demand for farm products increases far less rapidly than the demand for goods and services in general.

The second reason for lagging demand relates to population growth. Once a certain income level has been reached, each consumer's intake of food and fiber becomes relatively fixed. Thus subsequent increases in demand depend directly on growth in the number of consumers. In most advanced nations, including the United States, the demand for farm products increases at a rate roughly equal to the rate of population growth. Because U.S. population growth has not been rapid, the increase in U.S. demand for farm products has not kept pace with the rapid growth of farm output.

Graphical Portrayal

The combination of an inelastic and slowly increasing demand for agricultural products with a rapidly increasing supply puts strong downward pressure on agricultural prices and income. Figure 20.5 shows a large increase in agricultural supply accompanied by a very modest increase in demand. Because of the inelasticity of demand, those shifts result in a sharp decline in agricultural prices, accompanied by a relatively small increase in output. So farm income declines. On the graph, we see that farm

FIGURE 20.5 The long-run decline of agricultural prices and farm income. In the long run, increases in the demand for U.S. farm products (from D_1 to D_2) have not kept pace with the increases in supply (from S_1 to S_2) resulting from technological advances. Because agricultural demand is inelastic, these shifts have tended to depress agricultural prices (from P_1 to P_2) and reduce farm income (from $0P_1aQ_1$ to $0P_2bQ_2$) while increasing output only modestly (from Q_1 to Q_2).



income before the increases in demand and supply (measured by rectangle $0P_1aQ_1$) exceeds farm income after those increases ($0P_2bQ_2$). Because farm products have inelastic demand, an increase in supply relative to demand creates persistent downward pressure on farm income.

Consequences

The actual consequences of the demand and supply changes over time have been those predicted by the pure-competition model. The supply and demand conditions just outlined have increased the minimum efficient scale (MES) in agriculture and reduced crop prices. Farms that are too small to realize productivity gains and take advantage of economies of scale have discovered that their average total costs exceed the (declining) prices for their crops. So they can no longer operate profitably. In the long run, financial losses in agriculture have triggered a massive exit of workers to other sectors of the economy, as shown by Table 20.1. They have also caused a major consolidation of smaller farms into larger ones. A person farming, say, 240 acres of corn three decades ago is today likely to be farming two or three times that number of acres. Large corporate firms, collectively called **agribusiness,** have emerged in some areas of farming such as potatoes, beef, fruits, vegetables, and poultry. Today, there are 2.2 million farms compared to about 4 million in 1960, and farm labor constitutes about 1.1 percent of the U.S. labor force compared to 9.4 percent in 1960. (Global Perspective 20.1 compares the most recent labor-force percentages for several nations.)

TABLE 20.1 U.S. Farm Employment and Number of Farms, 1950–2009

	Farm Employment*		
Year	In Millions of People	As Percentage of Total Employment	Number of Farms, Thousands
1950	9.3	15.8	5,388
1960	6.2	9.4	3,962
1970	4.0	5.0	2,954
1980	3.5	3.5	2,440
1990	2.5	2.1	2,146
2000	2.2	1.6	2,172
2009	1.8	1.1	2,200

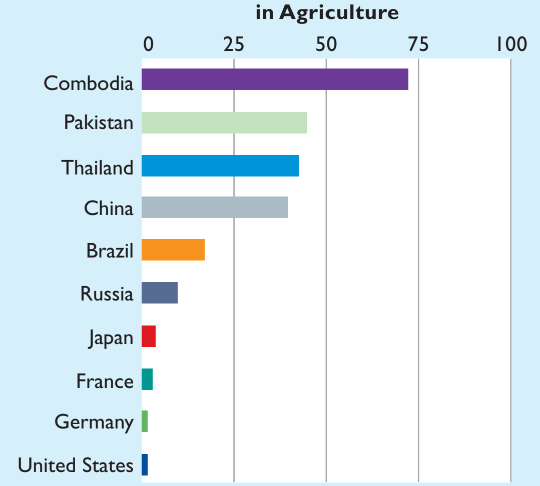
*Includes self-employed farmers, unpaid farmworkers, and hired farmworkers.

Sources: Derived by the authors from *Economic Report of the President, 2012,* Table B-100; U.S. Bureau of Labor Statistics, **www.bls.gov**; and Department of Agriculture, Economic Research Service, **www.ers.usda.gov**.

GLOBAL PERSPECTIVE 20.1

Average Percentage of Labor Force in Agriculture, Selected Nations, 2008–2012 Data

High-income nations devote a much smaller percentage of their labor forces to agriculture than do low-income nations. Because their workforces are so heavily committed to producing the food and fiber needed for their populations, low-income nations have relatively less labor available to produce housing, schools, autos, and the other goods and services that contribute to a high standard of living.



Source: World Development Indicators 2012, World Bank, **databank.worldbank.org**.

Farm-Household Income

Traditionally, the income of farm households was well below that of nonfarm households. But even with the lower real crop prices, that imbalance has reversed. In 2011—a particularly good year for agriculture—the average income of farm households was $87,289, compared to $69,677 for all U.S. households. Outmigration, consolidation, rising farm productivity, and significant government subsidies have boosted farm income *per farm household* (of which there are fewer than before).

Also, members of farm households operating smaller farms have increasingly taken jobs in nearby towns and cities. On average, only about 16 percent of the income of farm households derives from farming activities. This average, however, is pulled downward by the many households living in rural areas and operating small "residential farms." For households operating "commercial farms"—farms with annual sales of $250,000 or more—about 76 percent of the average income of $205,215 in 2011 derived from farming. Although agriculture is a declining industry, the 10 percent of farm households operating commercial farms in the United States are doing remarkably well, at least as a group.

QUICK REVIEW 20.1

- Agricultural prices and incomes are volatile in the short run because an inelastic demand converts small changes in farm output and demand into relatively larger changes in prices and income.
- Technological progress has generated large increases in the supply of farm products over time.
- Increases in demand for farm products have been modest in the United States because demand is inelastic with respect to income and because population growth has been modest.
- The combination of large increases in supply and small increases in demand has made U.S. agriculture a declining industry (as measured by the value of agricultural output as a percentage of GDP).

Economics of Farm Policy

LO20.3 Relate the rationale for farm subsidies and the economics and politics of price supports (price floors).

The U.S. government has subsidized agriculture since the 1930s with a "farm program" that includes (1) support for agricultural prices, income, and output; (2) soil and water conservation; (3) agricultural research; (4) farm credit; (5) crop insurance; and (6) subsidized sale of farm products in world markets.

We will focus on the main element of farm policy: the programs designed to prop up prices and income. This topic is particularly timely because in recent years (specifically, 1996, 2002, and 2008) Congress passed new farm laws replacing traditional forms of farm subsidies with new forms. To understand these new policies, we need to understand the policies they replaced and the purposes and outcomes of farm subsidies. Between 2002 and 2011, American farmers received an average of $21.7 billion of direct government subsidies each year. (As indicated in Global Perspective 20.2, farm subsidies are common in many nations.)

Rationale for Farm Subsidies

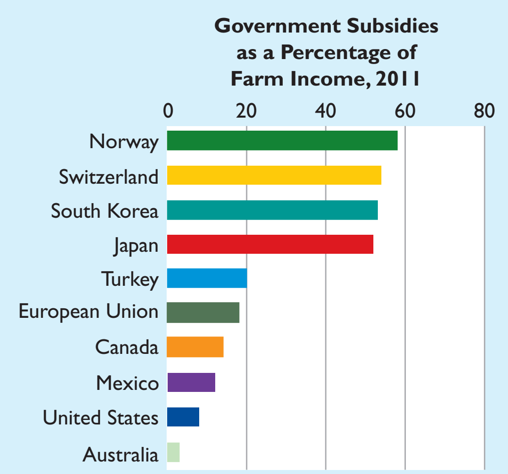
A variety of arguments have been made to justify farm subsidies over the decades:

- Although farm products are necessities of life, many farmers have relatively low incomes, so they should receive higher prices and incomes through public help.
- The "family farm" is a fundamental U.S. institution and should be nurtured as a way of life.
- Farmers are subject to extraordinary hazards—floods, droughts, and insects—that most other industries do

GLOBAL PERSPECTIVE 20.2

Agricultural Subsidies, Selected Nations

Farmers in various countries receive large percentages of their incomes as government subsidies.



Source: Producer and Consumer Support Estimates, Organization for Economic Cooperation and Development, **www.oecd.org/agriculture/pse**.

not face. Without government help, farmers cannot fully insure themselves against these disasters.

- While many farmers face purely competitive markets for their outputs, they buy inputs of fertilizer, farm machinery, and gasoline from industries that have considerable market power. Whereas those resource-supplying industries are able to control their prices, farmers are at the "mercy of the market" in selling their output. The supporters of subsidies argue that agriculture warrants public aid to offset the disadvantageous market-power imbalances faced by farmers.

Background: The Parity Concept

The Agricultural Adjustment Act of 1933 established the **parity concept** as a cornerstone of agricultural policy. The rationale of the parity concept can be stated in both real and nominal terms. In real terms, parity says that year after year for a fixed output of farm products, a farmer should be able to acquire a specific total amount of other goods and services. A particular real output should always result in the same real income: "If a farmer could take a bushel of corn to town in 1912 and sell it for enough money to buy a shirt, he should be able to sell a bushel of corn today and buy a shirt." In nominal terms, the parity concept suggests that the relationship between the prices received by farmers for their output and the prices they must pay for goods and services should remain constant. The parity concept implies that if the price of shirts tripled over some time period, then the price of corn should have tripled too. Such a situation is said to represent 100 percent of parity.

The **parity ratio** is the ratio of prices received to prices paid, expressed as a percentage. That is:

$$\text{Parity ratio} = \frac{\text{prices received by farmers}}{\text{prices paid by farmers}}$$

Why farmers would benefit from having the prices of their products based on 100 percent of parity is obvious. By 2010 nominal prices paid by farmers had increased 28-fold since 1900–1914, whereas nominal prices received by farmers had increased only about 9-fold. In 2010 the parity ratio stood at 0.42 (or 42 percent), indicating that prices received in 2012 could buy 42 percent as much as prices received in the 1910–1914 period. So a farm policy that enforced 100 percent of parity would generate substantially higher prices for farmers.

Economics of Price Supports

The concept of parity provides the rationale for government price floors on farm products. In agriculture those minimum prices are called **price supports.** We have shown that, in the long run, the market prices received by farmers have not kept up with the prices paid by them. One way to achieve parity, or some percentage thereof, is to have the government establish above-equilibrium price supports for farm products.

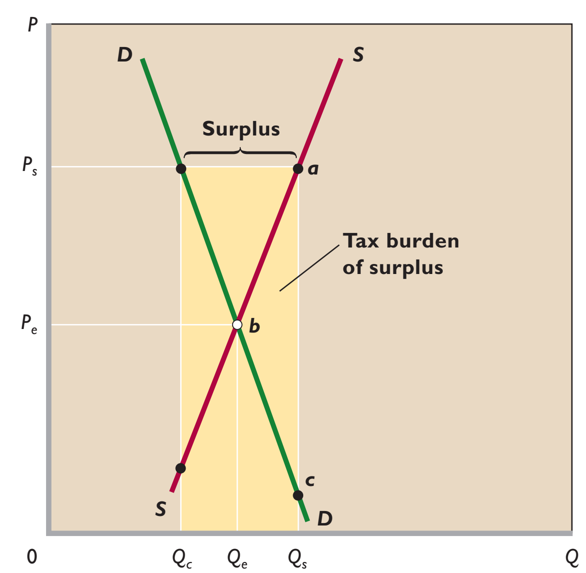
Many different price-support programs have been tried, but they all tend to have similar effects, some of which are subtle and negative. Suppose in Figure 20.6 that the equilibrium price is P_e and the price support is P_s. Then the major effects would be as follows.

Surplus Output The most obvious result is a product surplus. Consumers are willing to purchase only Q_c units at the supported price, while farmers supply Q_s units. What about the $Q_s - Q_c$ surplus that results? The government must buy it to make the above-equilibrium price support effective. As you will see, this surplus farm output means that agriculture receives an overallocation of resources.

Gain to Farmers Farmers benefit from price supports. In Figure 20.6, gross farm revenue rises from the free-market level represented by area $0P_ebQ_e$ to the larger, supported level shown by area $0P_saQ_s$.

Loss to Consumers Consumers lose; they pay a higher price (P_s rather than P_e) and consume less (Q_c rather than Q_e)

FIGURE 20.6 Price supports, agricultural surpluses, and transfers to farmers. The market demand D and supply S of a farm product yield equilibrium price P_e and quantity Q_e. An above-equilibrium price support P_s results in consumption of quantity Q_c, production of quantity Q_s, and a surplus of quantity $Q_s - Q_c$. The yellow rectangle represents a transfer of money from taxpayers to farmers. Triangle *bac* within the yellow rectangle shows the efficiency loss (or a deadweight loss) to society.



of the product. In some instances differences between the market price and the supported price are substantial. For example, the U.S.-supported price of a pound of sugar is about 32 percent higher than the world market price, and a quart of fluid milk is estimated to cost consumers twice as much as it would without government programs. Moreover, the burden of higher food prices falls disproportionately on the poor because they spend a larger part of their incomes on food.

Efficiency Losses Society loses because price supports create allocative inefficiency by encouraging an overallocation of resources to agriculture. A price floor (P_s) attracts more resources to the agricultural sector than would the free-market price (P_e). Viewed through the pure-competition model, the market supply curve in Figure 20.6 represents the marginal costs of all farmers producing this product at the various output levels. An efficient allocation of resources occurs at point *b*, where the market price P_e is equal to marginal cost. So the output Q_e reflects that efficient allocation of resources.

In contrast, the output Q_s associated with the price support P_s represents an overallocation of resources; for all units of output between Q_e and Q_s, marginal costs (measured on curve *S*) exceed the prices people are willing to pay for those units (measured on curve *D*). Simply stated, the marginal cost of the extra production exceeds its marginal benefit to society. Society incurs an efficiency loss (or a deadweight loss) of area *bac* because of the price-support system.

Other Social Losses Society at large loses in other ways. Taxpayers pay higher taxes to finance the government's purchase of the surplus. This added tax burden is equal to the surplus output $Q_s - Q_c$ multiplied by its price P_s, as shown by the yellow area in Figure 20.6. Recall, too, that the mere collection of taxes imposes an efficiency loss (Figure 18.10). Also, the cost of storing surplus farm output adds to this tax burden.

Government's intervention in agriculture also entails administrative costs. Thousands of government workers are needed to administer U.S. price supports and other farm programs.

Finally, the rent-seeking activity involved—the pursuit of political support to maintain price supports—is costly and socially wasteful. Farm groups spend considerable sums to sustain political support for price floors and other programs that enhance farm incomes.

Environmental Costs We know from Figure 20.6 that price supports encourage additional production. Although some of that extra output may come from the use of additional land, much of it comes from heavier use of fertilizer and pesticides. Those pesticides and fertilizers may pollute the environment (for example, groundwater) and create residues in food that pose health risks to farmworkers and consumers. Research shows a positive relationship between the level of price-support subsidies and the use of agrochemicals.

Farm policy also may cause environmental problems in less obvious ways. Farmers benefit from price supports only when they use their land consistently for a specific crop such as corn or wheat. That creates a disincentive to practice crop rotation, which is a nonchemical technique for controlling pests. Farm policy thus encourages the substitution of chemicals for other forms of pest control.

Also, we know from the concept of derived demand that an increase in the price of a product will increase the demand for relevant inputs. In particular, price supports for farm products increase the demand for land. And the land that farmers bring into farm production is often environmentally sensitive "marginal" land, such as steeply sloped, erosion-prone land, or wetlands that provide wildlife habitat. Similarly, price supports result in the use of more water for irrigation, and the resulting runoff may contribute to soil erosion.

International Costs Actually, the costs of farm price supports go beyond those indicated by Figure 20.6. Price supports generate economic distortions that cross national boundaries. For example, the high prices caused by price supports make the U.S. agricultural market attractive to foreign producers. But inflows of foreign agricultural products would serve to increase supplies in the United States, aggravating the problem of U.S. surpluses. To prevent that from happening, the United States is likely to impose import barriers in the form of tariffs or quotas. Those barriers tend to restrict the output of more-efficient foreign producers while encouraging more output from less-efficient U.S. producers. The result is a less-efficient use of world agricultural resources. This chapter's Last Word suggests that this is indeed the case for sugar.

Similarly, as the United States and other industrially advanced countries with similar agricultural programs dump surplus farm products on world markets, the prices of such products are depressed. Developing countries are often heavily dependent on world commodity markets for their incomes. So they are particularly hurt because their export earnings are reduced. Thus, U.S. subsidies for rice production have imposed significant costs on Thailand, a major rice exporter. Similarly, U.S. cotton programs have adversely affected Egypt, Mexico, and other cotton-exporting nations.

Reduction of Surpluses

Figure 20.6 suggests that programs designed to reduce market supply (shift S leftward) or increase market demand (shift D rightward) would help boost the market price toward the supported price P_s. Further, such programs would reduce or eliminate farm surpluses. The U.S. government has tried both supply and demand approaches to reduce or eliminate surpluses.

Restricting Supply Until recently, public policy focused mainly on restricting farm output. In particular, **acreage allotments** accompanied price supports. In return for guaranteed prices for their crops, farmers had to agree to limit the number of acres they planted in that crop. The U.S. Department of Agriculture first set the price support and then estimated the amount of the product consumers would buy at the supported price. It then translated that amount into the total number of planted acres necessary to provide it. The total acreage was apportioned among states, counties, and ultimately individual farmers.

These supply-restricting programs were only partially successful. They did not eliminate surpluses, mainly because acreage reduction did not result in a proportionate decline in production. Some farmers retired their worst land and kept their best land in production. They also cultivated their tilled acres more intensively. Superior seed, more and better fertilizer and insecticides, and improved farm equipment were used to enhance output per acre. And nonparticipating farmers expanded their planted acreage in anticipation of overall higher prices. Nevertheless, the net effect of acreage allotment undoubtedly was a reduction of farm surpluses and their associated costs to taxpayers.

Bolstering Demand Government has tried several ways to increase demand for U.S. agricultural products. For example, both government and private industry have spent large sums on research to create new uses for agricultural goods. The production of "gasohol," which is a blend of gasoline and alcohol (ethanol) made mainly from corn, is one such successful attempt to increase the demand for farm output. (See the nearby Consider This box for a fuller discussion of ethanol.) Recent attempts to promote "biodiesel," a fuel made from soybean oil and other natural vegetable oils, also fit the demand-enhancement approach.

The government has also created a variety of programs to stimulate consumption of farm products. For example, the objective of the food-stamp program is not only to reduce hunger but also to bolster the demand for food. Similarly, the Food for Peace program has enabled developing countries to buy U.S. surplus farm products with their own currencies, rather than having to use dollars. The federal government spends millions of dollars each year to advertise and promote global sales of U.S. farm products. Furthermore, U.S. negotiators have pressed hard in international trade negotiations to persuade foreign nations to reduce trade barriers to the importing of farm products.

CONSIDER THIS . . .

Putting Corn in Your Gas Tank

Government's promotion of greater production and use of corn-based ethanol serves both as a good example of an attempt by government to bolster the demand for U.S. farm products and as an example of how price changes can ripple through markets and produce myriad secondary effects. Gasoline producers blend ethanol (an alcohol-like substance) with conventional gasoline refined from oil. The government's rationale for promoting ethanol is to reduce U.S. dependency on foreign oil, but the strongest proponents are from states in the Corn Belt.

The ethanol program has several facets, including tariffs on imported ethanol, subsidies to oil refineries that buy ethanol, and mandates to industry to increase their use of alternative fuels. The rising demand for ethanol that resulted contributed to tripling of the inflation-adjusted price of a bushel of corn between 2005 and 2012.

But numerous secondary effects from the increased price of corn also occurred. Farmers shifted production toward corn and away from soybeans, sorghum, and other crops. The decreases in the supply of these other crops raised their prices, too. Also, because corn is used as a major feedstock, the price of beef, pork, and chicken rose.

The ethanol subsidies had other secondary effects. The prices of seed, fertilizer, and farmland all increased. Because corn is a water-intensive crop, its expanded production resulted in faster withdrawals of irrigation water from underground aquifers. The refining of ethanol also depleted ground water or removed it from rivers. Moreover, the increased use of fertilizer in corn production increased the runoff of nitrogen from fertilizer into streams and rivers, causing environmental damage.

The price effects of ethanol subsidies, however, may moderate as farmers shift additional land to corn, increasing its supply and reducing its price. Nevertheless, the multiple impacts of public policy illustrate an important economic maxim: In the economy, it is difficult to do just *one* thing.

During the era of price supports, the government's supply-restricting and demand-increasing efforts boosted agricultural prices and reduced surplus production, but they did not succeed in eliminating the sizable surpluses.

QUICK REVIEW 20.2

- The parity concept suggests that farmers should obtain a constant ratio of the prices they receive for their farm products and the prices they pay for goods and services in general.
- Price supports are government-imposed price floors (minimum prices) on selected farm products.
- Price supports cause surplus production (which the government must buy and store); raise farm income; increase food prices to consumers; and overallocate resources to agriculture.
- Domestic price supports encourage nations to erect trade barriers against imported farm products and to dump surplus farm products on world markets.

Criticisms and Politics

LO20.4 Describe major criticisms of the price-support system in agriculture.

After decades of experience with government price-support programs, it became apparent in the 1990s that farm policy was not working well. Major criticisms of farm subsidies emerged, as did a more skeptical analysis of the politics of those subsidies.

Criticisms of the Parity Concept

Economists uniformly rejected the rationale of the parity concept. They found no economic logic in the proposition that if a bushel of wheat could buy a shirt in 1900, it should still be able to buy a shirt several decades later. The relative values of goods and services are established by supply and demand, and those relative values change over time as technology changes, resource prices change, tastes change, and substitute resources and new products emerge. A fully equipped personal computer, monitor, and printer cost as much as a cheap new automobile in 1985. That was not true just a decade later because the price of computer equipment had dropped so dramatically. Based on the parity concept, one could argue that price supports and subsidies were justified for computer manufacturers!

Criticisms of the Price-Support System

Criticisms of the price-support system were equally severe.

Symptoms, Not Causes The price-support strategy in agriculture was designed to treat the symptoms, not the causes of the farm problem. The root cause of the long-run farm problem was misallocation of resources between agriculture and the rest of the economy. Historically, the problem had been one of too many farmers. The effect of that misallocation was relatively low agricultural prices and low farm income. But the price and income supports encouraged people to stay in farming rather than move to nonfarm occupations. That is, the price and income orientation of the farm program slowed the reallocation of resources necessary to resolve the long-run farm problem.

Misguided Subsidies Because price supports were on a per-bushel basis, the subsidy system benefited those farmers who needed subsidies the least. If the goal of farm policy was to raise low farm incomes, it followed that any program of federal aid should have been aimed at farmers with the lowest incomes. But the poor, low-output farmer did not produce and sell enough in the market to get much aid from price supports. Instead, the large, prosperous farmer reaped the benefits because of sizable output. On equity grounds, direct income payments to struggling farmers are highly preferable to indirect price-support subsidies that go primarily to large-scale, prosperous farmers. Better yet, say many economists, would be transition and retraining support for farmers willing to move out of farming and into other occupations and businesses in greater demand.

A related point concerns land values. The price and income benefits that the price-support system provided increased the value of farmland. By making crops more valuable, price supports made the land itself more valuable. That was helpful to farmers who owned the land they farmed but not to farmers who rented land. Farmers rented about 40 percent of their farmland, mostly from well-to-do nonfarm landlords. So, price supports became a subsidy to people who were not actively engaged in farming.

Policy Contradictions Because farm policy had many objectives, it often led to contradictions. Whereas most subsidized research was aimed at increasing farm productivity and the supply of farm products, acreage-allotment programs required that farmers take land out of production in order to reduce supply. Price supports for crops meant increased feed costs for ranchers and farmers and high consumer prices for animal products. Tobacco farmers were subsidized even though tobacco consumption was causing serious health problems. The U.S. sugar program raised prices for domestic producers by imposing import quotas that conflicted with free-trade policies. Conservation programs called for setting aside land for wildlife habitat, while price supports provided incentives to bring such acreage into production.

All these criticisms helped spawn policy reform. Nevertheless, as we will see, those reforms turned out to be less substantive than originally conceived. Nearly all these criticisms are as valid for current farm policy as they were for the price-support program.

The Politics of Farm Policy

In view of these criticisms, why did the United States continue its price-support program for 60 years and why does it still continue that program for sugar, milk, and tobacco? Why do farm subsidies in the billions of dollars still occur?

Public Choice Theory Revisited Public choice theory (Chapter 5) helps answer these questions. Recall that rent-seeking behavior occurs when a group (a labor union, firms in a specific industry, or farmers producing a particular crop) uses political means to transfer income or wealth to itself at the expense of another group or of society as a whole. And recall that the special-interest effect involves a program or policy from which a small group receives large benefits at the expense of a much larger group whose members individually suffer small losses. Both rent-seeking behavior and the special-interest effect help explain the politics of farm subsidies.

Suppose a certain group of farmers, say, peanut or sugar producers, organize and establish a well-financed political action committee (PAC). The PAC's job is to promote government programs that will transfer income to the group (this is rent-seeking behavior). The PAC vigorously lobbies U.S. senators and representatives to enact or to continue price supports, production quotas, or import quotas for peanuts or sugar. The PAC does this in part by making political contributions to sympathetic legislators. Although peanut production is heavily concentrated in a few states such as Georgia, Alabama, and Texas, the peanut PAC will also make contributions to legislators from other states in order to gain support.

But how can a small interest group like peanut or sugar growers successfully lobby to increase its own income at the expense of society as a whole? Because even though the total cost of the group's programs might be considerable, the cost imposed on each individual taxpayer is small (this is the special-interest effect). Taxpayers are likely to be uninformed about and indifferent to such programs since they have little at stake. Unless you grow sugar beets or peanuts, you probably have no idea how much these programs cost you as an individual taxpayer and consumer and therefore do not object when your legislator votes for, say, a sugar-support program. Thus, the PAC encounters little or no lobbying against its efforts.

Political logrolling—the trading of votes on policies and programs—also works to perpetuate certain programs: Senator Foghorn agrees to vote for a program that benefits Senator Moribund's constituents, and Moribund returns the favor. Example: Many members of Congress who represent low-income urban areas vote in favor of farm subsidies. In return, representatives of agricultural areas support such programs as food stamps, which subsidize food for the poor. The result is a rural-urban coalition through which representatives from both areas provide benefits for their constituents and enhance their reelection chances. Such coalitions help explain why farm subsidies persist and why the food-stamp program has been expanded over the years.

Large agribusinesses that supply inputs to agriculture also lend political support to farm subsidies because subsidies increase the amounts of agrochemicals and farm machinery that farmers are able to buy. And most of the thousands of government employees whose jobs depend on farm programs are highly supportive. So, too, are owners of farmland.

Public choice theory also tells us that politicians are likely to favor programs that have hidden costs. As we have seen, that is often true of farm programs. Our discussion of Figure 20.6 indicated that price supports involve not simply a transfer of money from taxpayer to farmer but costs that are hidden as higher food prices, storage costs for surplus output, costs of administering farm programs, and costs associated with both domestic and international misallocations of resources. Because those costs are largely indirect and hidden, farm programs are much more acceptable to politicians and the public than they would be if all costs were explicit.

Changing Politics In spite of rent seeking, special interests, and logrolling, a combination of factors has somewhat altered the politics of farm subsidies in recent decades.

Declining Political Support As the farm population declines, agriculture's political power weakens. The farm population was about 25 percent of the general population in the 1930s, when many U.S. farm programs were established; now it is less than 2 percent. Urban congressional representatives now constitute a 10-to-1 majority over their rural colleagues. An increasing number of legislators are critically examining farm programs for their effects on consumers' grocery bills as well as on farm incomes. Also, more farmers themselves are coming to resent the intrusion of the federal government into their farming decisions. A few rural-state congressional members now support free-market agriculture.

World Trade Considerations The United States has taken the lead to reduce barriers to world trade in agricultural products. This has also contributed to the more critical attitude toward farm subsidies, particularly price supports. The nations of the European Union (EU) and many other nations support agricultural prices. And, to maintain their high domestic prices, they restrict imports of foreign farm products by imposing tariffs and quotas. They then try to rid themselves of their domestic surpluses by subsidizing exports into world markets. The effects on the United States are that (1) trade barriers hinder U.S. farmers from selling to EU nations and (2) subsidized exports from those nations depress world prices for agricultural products, making world markets less attractive to U.S. farmers.

Perhaps most important, farm programs such as those maintained by the EU and the United States distort both world agricultural trade and the international allocation of agricultural resources. Encouraged by artificially high prices, farmers in industrially advanced nations produce more food and fiber than they would otherwise. The resulting surpluses flow into world markets, where they depress prices. This means that farmers in countries with no farm programs—many of them developing countries—face artificially low prices for their exports, which signals them to produce less. Overall, the result is a shift in production away from what would occur on the basis of comparative advantage. As an example, price supports cause U.S. agricultural resources to be used for sugar production, even though sugar can be produced at perhaps half the cost in the Caribbean countries and Australia.

Recognizing these distortions, in 1994 the 128 nations then belonging to the World Trade Organization (WTO) agreed to reduce farm price-support programs by 20 percent by the year 2000 and to reduce tariffs and quotas on imported farm products by 15 percent. Larger, more significant, reductions of farm subsidies and agricultural tariffs are part of the agenda of the most recent round of trade negotiations (the Doha Development Agenda). But reaching agreement on those reductions has proved difficult. As of early 2013, negotiations over these issues were completely stalled.

Recent Farm Policies

LO20.5 List the main elements of existing federal farm policy.

In the mid-1990s there was a common feeling among economists and political leaders that the goals and techniques of farm policy needed to be reexamined and revised. Moreover, crop prices were relatively high at the time and Congress wanted to reduce large federal budget deficits.

Freedom to Farm Act of 1996

In 1996 Congress radically revamped 60 years of U.S. farm policy by passing the **Freedom to Farm Act.** The law ended price supports and acreage allotments for wheat, corn, barley, oats, sorghum, rye, cotton, and rice. Farmers were allowed to respond to changing crop prices by planting as much or as little of these crops as they chose. Also, they were free to plant crops of their choice. If the price of, say, oats increased, farmers could plant more oats and less barley. Markets, not government programs, were to determine the kinds and amounts of crops grown.

To ease the transition away from price supports, the Freedom to Farm Act granted declining annual transition payments through 2002. The $37 billion of total scheduled payments through 2002 was based on the production levels of the crops each farmer previously had grown under the price-support system. So a previous wheat farmer, for example, would receive cash payments for 7 years regardless of the current price of wheat or amount of wheat presently grown.

But this ambitious plan to wean American agriculture from subsidies unraveled in 1998 and 1999, when sharply reduced export demand and strong crop production in the United States depressed the prices of many farm products. Congress responded by supplementing the previously scheduled transition payments with large "emergency aid" payments to farmers. Agricultural subsidies for 1999–2002 averaged $20 billion annually—even more than they were before passage of the Freedom to Farm Act.

The Food, Conservation, and Energy Act of 2008

Since 2002, agricultural policy in the United States has substantially retreated from the free-market intent of the 1996 law. Current subsidy programs continue the "freedom to plant" approach and provide revenue guarantees for farmers by ironically transforming the "transition payments" policy into a permanent program known as "direct payments." These revenue guarantees kick in automatically when crop prices (or total revenues) fall below targeted levels.

The **Food, Conservation, and Energy Act of 2008** laid the foundation for the current system of agricultural subsides. The law provided three main forms of cash commodity subsidies, with an option on one of them.

Direct Payments The **direct payments** under the 2008 law are similar to the transition payments paid under the Freedom to Farm Act. The cash payments are fixed for each crop based on a farmer's historical pattern of production and are unaffected by current crop prices or current

LAST WORD

The Sugar Program: A Sweet Deal

The Sugar Program Is a Sweet Deal for Domestic Sugar Producers, but It Imposes Heavy Costs on Domestic Consumers, Domestic Candy Manufacturers, Foreign Producers, and the American Economy.

The continuing U.S. sugar program uses price supports and import quotas to guarantee a minimum price of sugar for domestic sugar producers. The program has significant effects, both domestically and internationally.

Domestic Costs Price supports and import quotas have boosted the domestic price of sugar to approximately 32 percent above the world price (for 2012, $0.29 per pound compared to the international price of $0.22 per pound). The aggregate cost to domestic consumers has been estimated at between $1.5 billion and $1.9 billion per year. In contrast, each sugar producer receives from subsidies alone an amount estimated to be twice the nation's average family income. In one particular year, a single producer received an estimated $30 million in benefits. Many sugar producers obtain more than $1 million each year in benefits.

Import Quotas As a consequence of high U.S. domestic price supports, foreign sugar producers have a strong incentive to sell their output in the United States. But an influx of lower-priced foreign sugar into the U.S. domestic market would undermine U.S. price supports. The government therefore has imposed import quotas on foreign sugar. It decides how much sugar can be imported at a zero or very low tariff rate, and then it charges a prohibitively high tariff for any quantities above that amount. As the gap between U.S.-supported prices and world prices has widened, imports have declined as a percentage of sugar consumed in the United States. In 1975, about 30 percent of the sugar consumed in the United States was imported; currently about 20 percent comes from abroad. Domestic policy regarding the U.S. sugar industry largely dictates the nation's international trade policy with respect to sugar.

Developing Countries The loss of the U.S. market has had several harmful effects on sugar-exporting developing countries such as the Philippines, Brazil, and several Central American countries.

First, exclusion from the U.S. market has significantly reduced their export revenues—by an amount estimated to be many billions of dollars per year. That decline in export revenues is important because many of the sugar-producing countries depend on such revenues to pay interest and principal on large debts owed to the United States and other industrially advanced nations.

production. Farmers are free to plant as much or as little of any particular crop as they want and still receive these payments. These direct payments do not decline from year to year. They are a permanent transfer payment from the federal government (general taxpayers) to farmers. The payments range from 2.4 cents per bushel for oats up to 54 cents per bushel for wheat. The rate for corn is 28 cents per bushel and the rate for soybeans is 44 cents per bushel.

Countercyclical Payments This component of farm policy ties a separate set of subsidies to the difference between market prices of specified farm products and a target price set for each crop. Like direct payments, these **countercyclical payments (CCPs)** are based on previous crops grown and are received regardless of the current crop planted. For example, the target price for corn in the years 2008–2012 is $2.63 per bushel. If corn is at or exceeds $2.63 in one of those years, the farmer who qualifies will receive no CCP. But if the price is below $2.63, the farmer will receive CCP payments geared to the size of the price gap. The CCP system has returned a form of price supports to a prominent role in farm policy, but it bases those supports on past crops grown, not current crops planted.

Beginning in 2009, farmers obtained the option of withdrawing from the CCP program and instead participating in the Average Crop Revenue Election (ACRE). This program bases farmers' countercyclical payments on average crop yields per acre in their states over the past five years along with the average national price for the crop in the past two years.

Marketing Loans Finally, current law contains a **marketing loan program** under which farmers can receive a loan (on a per-unit-of-output basis) from a government

Second, barred by quotas from sale in the U.S. market, the sugar produced by the developing countries has been added to world markets, where the increased supply has depressed the world price of sugar.

Third, domestic price supports have caused U.S. sugar production to expand to the extent that the United States may soon change from a sugar-importing to a sugar-exporting nation. That is, the U.S. sugar program may soon be a source of new competition for the sugar producers of the developing countries. Sugar price supports in the European Union have already turned that group of nations into sugar exporters.

U.S. Efficiency Loss The sugar program benefits sugar producers by about $1 billion annually but costs U.S. consumers about $1.5 billion to $1.9 billion each year. The excess of losses over gains is therefore $500 million to $900 million annually. This efficiency loss (or deadweight loss) results from the overallocation of U.S. resources to growing and processing sugar beets and sugar cane.

As a secondary effect, the higher domestic sugar prices have encouraged several U.S. confectionery firms (candy manufacturers) to relocate their operations to Canada or Mexico. According to the U.S. Commerce Department, for every American job that has been added by the price supports in the cane sugar and sugar beet industries, three American jobs have been lost in the industries buying sugar. Government economists estimate that the confectionery industry has lost a total of about 24,000 jobs since 2003.

Global Resource Misallocation
Both domestically and globally, the sugar price-support programs of the United States and other industrially advanced economies have distorted the worldwide allocation of agricultural resources. Price supports have caused a shift of resources to sugar production by less efficient U.S. producers, and U.S. import quotas and consequent low world sugar prices have caused more efficient foreign producers to restrict their production. Thus high-cost producers are producing more sugar and low-cost producers are producing less, resulting in the inefficient use of the world's agricultural resources.

Adding to the inefficient use of world resources, the relocation of candy manufacturers—to avoid artificially sweetened U.S. sugar prices—is moving capital and labor resources away from their place of comparative advantage.

lender. If the crop price at harvest is higher than the price specified in the loan (the loan price), farmers can repay their loans, with interest. If the crop price is lower than the loan price, farmers can forfeit their harvested crops to the lender and be free of their loans. In this second case, farmers receive what amounts to a subsidy because the proceeds from the loan exceed the revenues from the sale of the crop in the market.

The 2008 farm law reduces the risk of price and revenue variability for farmers and increases farm income. But the law fails to address the problem of subsidies. However structured, subsidies slow the exodus of resources from agriculture and maintain high production levels. This means lower crop prices and less market income for farmers. These lower prices and reduced market incomes, in turn, provide the rationale for continued government subsidies!

QUICK REVIEW 20.3

- Farm policy in the United States has been heavily criticized for delaying the shift of resources away from farming, directing most subsidies to wealthier farmers, and being fraught with policy contradictions.
- The persistence of farm subsidies can largely be explained in terms of rent-seeking behavior, the special-interest effect, political logrolling, and other aspects of public choice theory.
- The Freedom to Farm Act of 1996 eliminated price supports and acreage allotments for many of the nation's crops, while continuing direct subsidies to farmers.
- The Food, Conservation, and Energy Act of 2008 provides three major kinds of farm subsidies: direct payments, countercyclical payments, and marketing loans.

SUMMARY

LO20.1 Explain why agricultural prices and farm income are unstable.

In the short run, the highly inelastic demand for farm products transforms small changes in output and small shifts in demand into large changes in prices and income.

LO20.2 Discuss why there has been a huge employment exodus from agriculture to other U.S. industries over the past several decades.

Over the long run, rapid technological advance, together with a highly inelastic and relatively slow-growing demand for agricultural output, has made agriculture a declining industry in the United States and dictated that resources exit the industry.

LO20.3 Relate the rationale for farm subsidies and the economics and politics of price supports (price floors).

Historically, farm policy has been centered on price and based on the parity concept, which suggests that the relationship between prices received and paid by farmers should be constant over time.

The use of price floors or price supports has a number of economic effects: It (*a*) causes surplus production; (*b*) increases the incomes of farmers; (*c*) causes higher consumer prices for farm products; (*d*) creates an overallocation of resources to agriculture; (*e*) obliges society to pay higher taxes to finance the purchase and storage of surplus output; (*f*) increases pollution because of the greater use of agrochemicals and vulnerable land; and (*g*) forces other nations to bear the costs associated with import barriers and depressed world agricultural prices.

With only limited success, the federal government has pursued programs to reduce agricultural supply and increase agricultural demand as a way to reduce the surpluses associated with price supports.

LO20.4 Describe major criticisms of the price-support system in agriculture.

Economists have criticized U.S. farm policy for (*a*) confusing symptoms (low farm incomes) with causes (excess capacity), (*b*) providing the largest subsidies to high-income farmers, and (*c*) creating contradictions among specific farm programs.

The persistence of agricultural subsidies can be explained by public choice theory and, in particular, as rent-seeking behavior; the special-interest effect; and political logrolling.

Political backing for price supports and acreage allotments has eroded for several reasons: (*a*) The number of U.S. farmers, and thus their political clout, has declined relative to the number of urban consumers of farm products and (*b*) successful efforts by the United States to get other nations to reduce their farm subsidies have altered the domestic debate on the desirability of U.S. subsidies.

LO20.5 List the main elements of existing federal farm policy.

The Freedom to Farm Act of 1996 ended price supports and acreage allotments for wheat, corn, barley, oats, sorghum, rye, cotton, and rice. The law established declining annual transition payments through the year 2002, but those payments were no longer tied to crop prices or the current crop produced.

When crop prices plummeted in 1998 and 1999, Congress supplemented the transition payments of the Freedom to Farm Act with large amounts of emergency aid. Total subsidies to agriculture averaged $20 billion annually in the years 1999–2002.

Beginning in 2002, the federal government retreated from the free-market principles of the Freedom to Farm Act, setting up a system of permanent direct payments to farmers along with countercyclical farm-revenue guarantees.

The Food, Conservation, and Energy Act of 2008 provides farmers with direct payments (based on previous crops planted), countercyclical payments (based on the differences between market prices and targeted prices), and marketing loans (based on a specified crop price and an option to either pay back the loan or forfeit the crop to the government lender).

TERMS AND CONCEPTS

farm commodities
food products
agribusiness
parity concept
parity ratio
price supports
acreage allotments
Freedom to Farm Act
Food, Conservation, and Energy Act of 2008
direct payments
countercyclical payments (CCPs)
marketing loan program

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Carefully evaluate: "The supply and demand for agricultural products are such that small changes in agricultural supply result in drastic changes in prices. However, large changes in agricultural prices have modest effects on agricultural

output." (Hint: A brief review of the distinction between *supply* and *quantity supplied* may be helpful.) Do exports increase or reduce the instability of demand for farm products? Explain. **LO20.1**

2. What relationship, if any, can you detect between the facts that farmers' fixed costs of production are large and the supply of most agricultural products is generally inelastic? Be specific in your answer. **LO20.1**
3. Explain how each of the following contributes to the farm problem: **LO20.1, LO20.2**
   a. The inelasticity of demand for farm products.
   b. The rapid technological progress in farming.
   c. The modest long-run growth in demand for farm commodities.
   d. The volatility of export demand.
4. The key to efficient resource allocation is shifting resources from low-productivity to high-productivity uses. In view of the high and expanding physical productivity of agricultural resources, explain why many economists want to divert additional resources away from farming in order to achieve allocative efficiency. **LO20.2**
5. Explain and evaluate: "Industry complains of the higher taxes it must pay to finance subsidies to agriculture. Yet the trend of agricultural prices has been downward, while industrial prices have been moving upward, suggesting that on balance agriculture is actually subsidizing industry." **LO20.3**
6. "Because consumers as a group must ultimately pay the total income received by farmers, it makes no real difference whether the income is paid through free farm markets or through price supports supplemented by subsidies financed out of tax revenue." Do you agree? **LO20.3**
7. If in a given year the indexes of prices received and paid by farmers were 120 and 165, respectively, what would the parity ratio be? Explain the meaning of that ratio. **LO20.3**
8. Explain the economic effects of price supports. Explicitly include environmental and global impacts in your answer. On what grounds do economists contend that price supports cause a misallocation of resources? **LO20.3**
9. Do you agree with each of the following statements? Explain why or why not. **LO20.3, LO20.4**
   a. The problem with U.S. agriculture is that there are too many farmers. That is not the fault of farmers but the fault of government programs.
   b. The federal government ought to buy up all U.S. farm surpluses and give them away to developing nations.
   c. All industries would like government price supports if they could get them; agriculture has obtained price supports only because of its strong political clout.
10. What are the effects of farm subsidies such as those of the United States and the European Union on (*a*) domestic agricultural prices, (*b*) world agricultural prices, and (*c*) the international allocation of agricultural resources? **LO20.3**
11. Use public choice theory to explain the persistence of farm subsidies in the face of major criticisms of those subsidies. If the special-interest effect is so strong, what factors made it possible in 1996 for the government to end price supports and acreage allotments for several crops? **LO20.4**
12. What was the major intent of the Freedom to Farm Act of 1996? Do you agree with the intent? Why or why not? Did the law succeed in reducing overall farm subsidies? Why or why not? **LO20.5**
13. Distinguish the major features of direct subsidies, countercyclical payments, and marketing loan subsidies under the Food, Conservation, and Energy Act of 2008. In what way do countercyclical payments and marketing loans help reduce the volatility of farm income? In what way do direct subsidies perpetuate the long-run farm problem of too many resources in agriculture? **LO20.5**
14. **LAST WORD** What groups benefit and what groups lose from the U.S. sugar subsidy program?

REVIEW QUESTIONS

1. Suppose that the demand for olive oil is highly inelastic. Also suppose that the supply of olive oil is fixed for the year. If the demand for olive oil suddenly increases because of a shortage of corn oil, you would expect a ___________ in the price of olive oil. **LO20.2**
   a. Large increase.
   b. Small increase.
   c. Large decrease.
   d. Small decrease.
   e. No change.
2. Use supply and demand curves to depict equilibrium price and output in a competitive market for some farm product. Then show how an above-equilibrium price floor (price support) would cause a surplus in this market. Demonstrate in your graph how government could reduce the surplus through a policy that (*a*) changes supply or (*b*) changes demand. Identify each of the following actual government policies as primarily affecting the supply of or the demand for a particular farm product: acreage allotments, the food-stamp program, the Food for Peace program, a government buyout of dairy herds, and export promotion. **LO20.3**
3. Suppose that the government has been supporting the price of corn. Its free market price is \$2.50 per bushel, but the government has been setting a support price of \$3.50 per bushel. Which of the following are ways that the government might try to reduce the size of the corn surplus? **LO20.3**
   *Select **one or more** answers from the choices shown.*
   a. Decrease the support price.
   b. Institute an acreage allotment program.
   c. Decrease demand by taxing purchases of corn.
   d. Raise the support price.

4. The majority of farm subsidies flow toward ____________. **LO20.4**
   - **a.** Poor, small-scale farmers.
   - **b.** Rich, large-scale farmers.
   - **c.** Government employees.
   - **d.** Grain wholesalers.

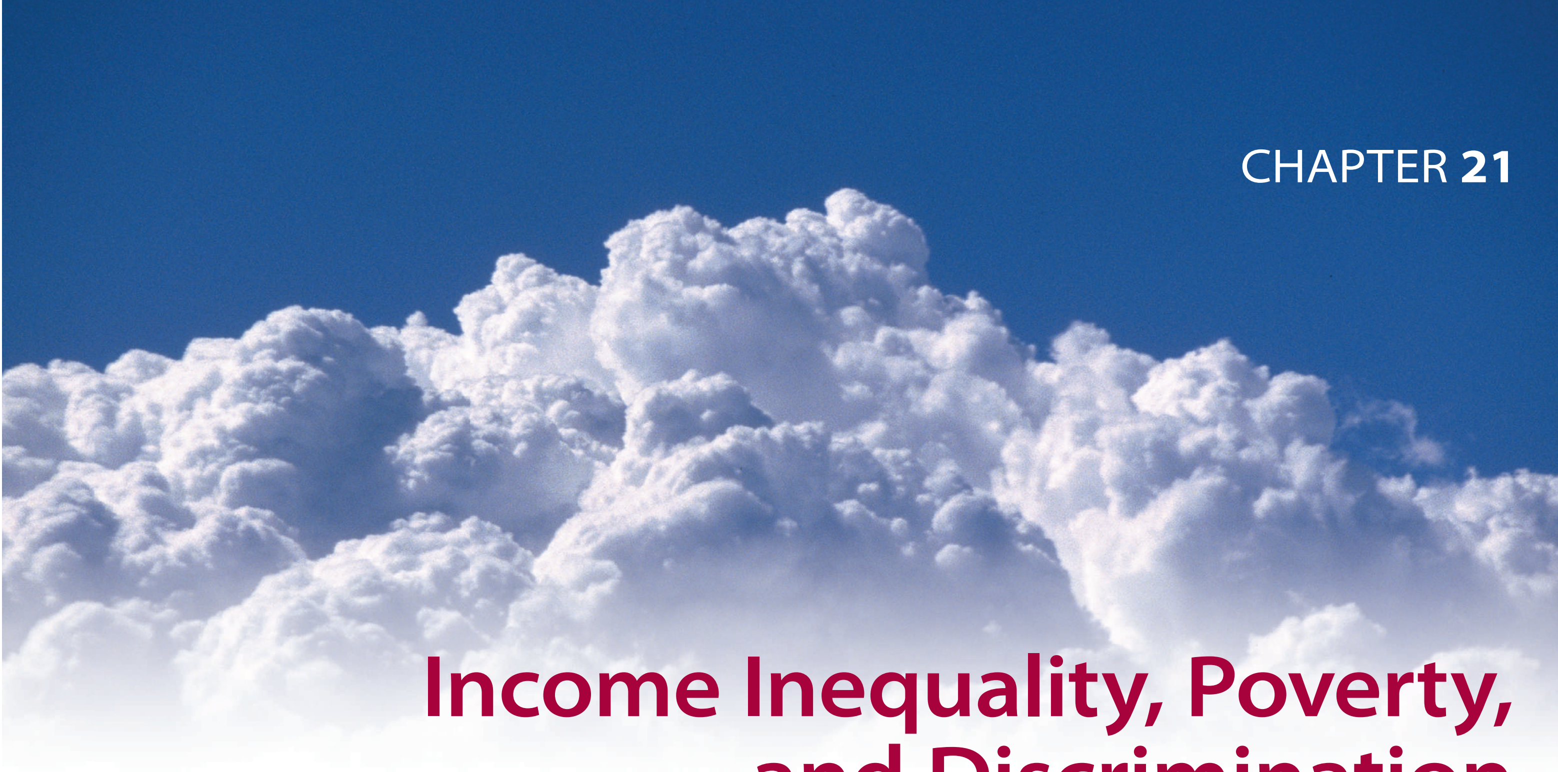
5. Which of the following are elements of current U.S. farm policy? **LO20.5**
   - **a.** Farmers are free to choose how much to plant of any particular crop.
   - **b.** Direct payments.
   - **c.** Price supports.
   - **d.** Countercyclical payments.

PROBLEMS

1. Suppose that corn currently costs \$4 per bushel and that wheat currently costs \$3 per bushel. Also assume that the price elasticity of corn is 0.10, while the price elasticity of wheat is 0.15. For the following questions about elasticities, simply use the percentage changes that are provided rather than attempting to calculate those percentage changes yourself using the midpoint formula given in Chapter 6. **LO20.1**
   - **a.** If the price of corn fell by 25 percent to \$3 per bushel, by what percentage would the quantity demanded of corn increase? What if the price of corn fell by 50 percent to \$2 per bushel?
   - **b.** To what value would the price of wheat have to fall to induce consumers to increase their purchases of wheat by 5 percent?
   - **c.** If the government imposes a \$0.40 per bushel tax on corn so that the price of corn rises by 10 percent to \$4.40 per bushel, by what percentage would the quantity demanded of corn decrease? If the initial quantity demanded is 10 billion bushels per year, by how many bushels would the quantity demanded decrease in response to this tax?
2. Suppose that both wheat and corn have an income elasticity of 0.1. **LO20.1**
   - **a.** If the average income in the economy increases by 2 percent each year, by what percentage does the quantity demanded of wheat increase each year, holding all other factors constant? Holding all other factors constant, if 10 billion bushels are demanded this year, by how many bushels will the quantity demanded increase next year if incomes rise by 2 percent?
   - **b.** Given that average personal income doubles in the United States about every 30 years, by about what percentage does the quantity demanded of corn increase every 30 years, holding all other factors constant?
3. Suppose that 10 workers were required in 2010 to produce 40,000 bushels of wheat on a 1,000-acre farm. **LO20.2**
   - **a.** What is the average output per acre? Per worker?
   - **b.** If in 2020 only 8 workers produce 44,000 bushels of wheat on that same 1,000-acre farm, what will be the average output per acre? Per worker?
   - **c.** By what percentage does productivity (output per worker) increase over those 10 years? Over those 10 years, what is the average annual percentage increase in productivity?
4. In 2009, it was estimated that the total value of all corn-production subsidies in the United States totaled about \$4 billion. The population of the United States was approximately 300 million people that year. **LO20.3**
   - **a.** On average, how much did corn subsidies cost per person in the United States in 2009? (Hint: A billion is a 1 followed by nine zeros. A million is a 1 followed by six zeros.)
   - **b.** If each person in the United States is only willing to spend \$0.50 to support efforts to overturn the corn subsidy, and if antisubsidy advocates can only raise funds from 10 percent of the population, how much money will they be able to raise for their lobbying efforts?
   - **c.** If the recipients of corn subsidies donate just one percent of the total amount that they receive in subsidies, how much could they raise to support lobbying efforts to continue the corn subsidy?
   - **d.** By how many dollars does the amount raised by the recipients of the corn subsidy exceed the amount raised by the opponents of the corn subsidy?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.



CHAPTER **21**

Income Inequality, Poverty, and Discrimination

Learning Objectives

LO21.1 **Explain how income inequality in the United States is measured and described.**

LO21.2 **Discuss the extent and sources of income inequality.**

LO21.3 **Demonstrate how income inequality has changed since 1975.**

LO21.4 **Debate the economic arguments for and against income inequality.**

LO21.5 **Relate how poverty is measured and its incidence by age, gender, ethnicity, and other characteristics.**

LO21.6 **Identify the major components of the income-maintenance program in the United States.**

LO21.7 **Discuss labor market discrimination and how it might affect hiring decisions and wages.**

Evidence that suggests wide income disparity in the United States is easy to find. In 2012 former talk-show host Oprah Winfrey earned \$175 million, *Transformers* director Michael Bay earned \$160 million, and singer Brittany Spears earned \$58 million. In contrast, the salary of the president of the United States is \$400,000, and the typical schoolteacher earns \$56,000. A full-time minimum-wage worker at a fast-food restaurant makes about \$15,000. Cash welfare payments to a mother with two children average \$5,150 per year.

In 2011 about 46.2 million Americans—or 15.0 percent of the population—lived in poverty. An estimated 636,000 people were homeless that year.

The richest fifth of American households received about 51.1 percent of total income, while the poorest fifth received about 3.2 percent.

What are the sources of income inequality? Is income inequality rising or falling? Is the United States making progress against poverty? What are the major income-maintenance programs in the United States? What role does discrimination play in reducing wages for some and increasing wages for others? These are some of the questions we will answer in this chapter.

Facts about Income Inequality

LO21.1 Explain how income inequality in the United States is measured and described.

Average household income in the United States is among the highest in the world; in 2011, it was $69,821 per household (one or more persons occupying a housing unit). But that average tells us nothing about income inequality. To learn about that, we must examine how income is distributed around the average.

Distribution by Income Category

One way to measure **income inequality** is to look at the percentages of households in a series of income categories. Table 21.1 shows that 25.0 percent of all households had annual before-tax incomes of less than $25,000 in 2011, while 21.0 percent had annual incomes of $100,000 or more. The data in the table suggest a wide dispersion of household income and considerable inequality of income in the United States.

Distribution by Quintiles (Fifths)

A second way to measure income inequality is to divide the total number of individuals, households, or families (two or more persons related by birth, marriage, or adoption) into five numerically equal groups, or *quintiles*, and examine the percentage of total personal (before-tax) income received by each quintile. We do this for households in the table in Figure 21.1, where we also provide the upper income limit for each quintile. Any amount of income greater than that listed in each row of column 3 would place a household into the next-higher quintile.

TABLE 21.1 **The Distribution of U.S. Income by Households, 2009**

(1) Personal Income Category	(2) Percentage of All Households in This Category
Under $15,000	13.5
$15,000–$24,999	11.5
$25,000–$34,999	10.9
$35,000–$49,999	13.9
$50,000–$74,999	17.6
$75,000–$99,999	11.5
$100,000 and above	21.0
	100.0

Source: Bureau of the Census, **www.census.gov**. Numbers do not add to 100 percent due to rounding.

The Lorenz Curve and Gini Ratio

We can display the quintile distribution of personal income through a **Lorenz curve.** In Figure 21.1, we plot the cumulative percentage of households on the horizontal axis and the percentage of income they obtain on the vertical axis. The diagonal line 0*e* represents a *perfectly equal distribution of income* because each point along that line indicates that a particular percentage of households receive the same percentage of income. In other words, points representing 20 percent of all households receiving 20 percent of total income, 40 percent receiving 40 percent, 60 percent receiving 60 percent, and so on, all lie on the diagonal line.

By plotting the quintile data from the table in Figure 21.1, we obtain the Lorenz curve for 2011. The bottom 20 percent of all households received 3.2 percent of the income, as shown by point *a*; the bottom 40 percent received 11.6 percent (= 3.2 + 8.4), as shown by point *b*; and so forth. The blue area between the diagonal line and the Lorenz curve is determined by the extent that the Lorenz curve sags away from the diagonal and indicates the degree of income inequality. If the actual income distribution were perfectly equal, the Lorenz curve and the diagonal would coincide and the blue area would disappear.

WORKED PROBLEMS

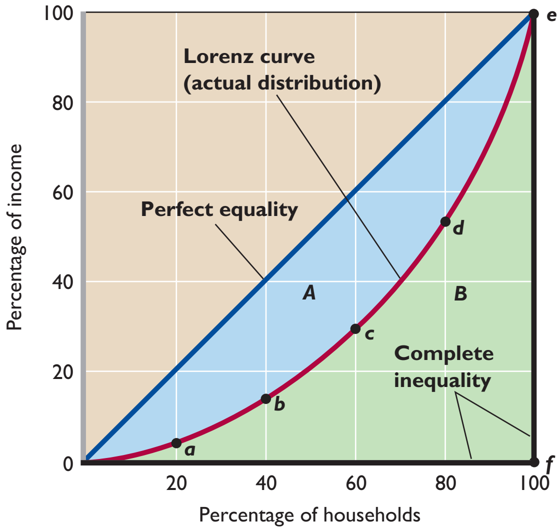
W21.1
Lorenz curve

At the opposite extreme is complete inequality, where all households but one have zero income. In that case the Lorenz curve would coincide with the horizontal axis

FIGURE 21.1 The Lorenz curve and Gini ratio. The Lorenz curve is a convenient way to show the degree of income inequality (here, household income by quintile in 2011). The area between the diagonal (the line of perfect equality) and the Lorenz curve represents the degree of inequality in the distribution of total income. This inequality is measured numerically by the Gini ratio—area *A* (shown in blue) divided by area *A* + *B* (the blue + green area). The Gini ratio for the distribution shown is 0.477.

(1) Quintile (2011)	(2) Percentage of Total Income	(3) Upper Income Limit
Lowest 20 percent	3.2	$ 20,262
Second 20 percent	8.4	38,520
Third 20 percent	14.3	62,434
Fourth 20 percent	23.0	101,582
Highest 20 percent	51.1	No limit
Total	100.0	

Source: Bureau of the Census, **www.census.gov**.



from 0 to point *f* (at 0 percent of income) and then would move immediately up from *f* to point *e* along the vertical axis (indicating that a single household has 100 percent of the total income). The entire area below the diagonal line (triangle 0*ef*) would indicate this extreme degree of inequality. So the farther the Lorenz curve sags away from the diagonal, the greater is the degree of income inequality.

The income inequality described by the Lorenz curve can be transformed into a **Gini ratio**—a numerical measure of the overall dispersion of income:

$$\text{Gini ratio} = \frac{\text{area between Lorenz curve and diagonal}}{\text{total area below the diagonal}}$$

$$= \frac{A \text{ (blue are)}}{A + B \text{ (blue + green area)}}$$

The Gini ratio is 0.477 for the distribution of household income shown in Figure 21.1. As the area between the Lorenz curve and the diagonal gets larger, the Gini ratio rises to reflect greater inequality. Lower Gini ratios denote less inequality; higher ratios indicate more inequality. The Gini coefficient for complete income equality is zero and for complete inequality is 1.

Because Gini ratios are numerical, they are easier to use than Lorenz curves for comparing the income distributions of different ethnic groups and countries. For example, in 2011 the Gini ratio of U.S. household income for Hispanics was 0.458; for whites, 0.469; for Asians, 0.466; and for African-Americans, 0.502.[1] Gini ratios of various nations range from 0.230 (Sweden) to 0.707 (Namibia). Examples within this range include Denmark, 0.248; Italy, 0.319; Mexico, 0.517; and South Africa, 0.650.[2]

Income Mobility: The Time Dimension

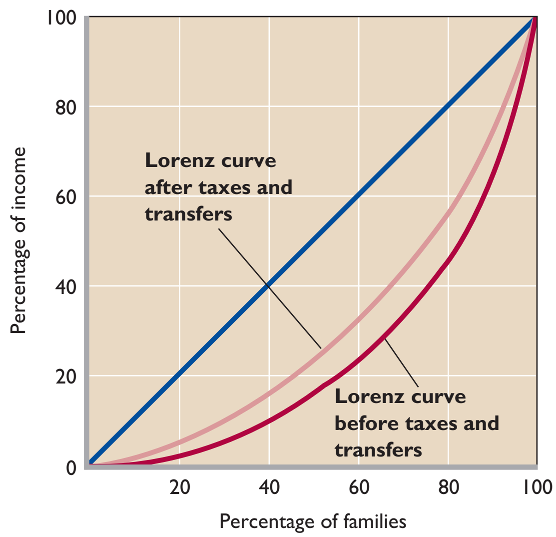
The income data used so far have a major limitation: The income accounting period of 1 year is too short to be very meaningful. Because the Census Bureau data portray the distribution of income in only a single year, they may conceal a more equal distribution over a few years, a decade, or even a lifetime. If Brad earns $1,000 in year 1 and $100,000 in year 2, while Jenny earns $100,000 in year 1 and only $1,000 in year 2, do we have income inequality? The answer depends on the period of measurement. Annual data would reveal great income inequality, but there would be complete equality over the 2-year period.

This point is important because evidence suggests considerable "churning around" in the distribution of income over time. Such movement of individuals or households from one income quintile to another over time is called **income mobility.** For most income receivers, income starts at a relatively low level during youth, reaches a peak during middle age, and then declines. It follows that

[1]U.S. Census Bureau, *Historical Income Tables*, **www.census.gov**.

[2]*CIA World Factbook*, *2012*, **www.cia.gov**.

FIGURE 21.2 The impact of taxes and transfers on U.S. income inequality. The distribution of household income is significantly more equal after taxes and transfers are taken into account than before. Transfers account for most of the lessening of inequality and provide most of the income received by the lowest quintile of households.



Quintile	Percentage of Total Income Received, 2009 (1) Before Taxes and Transfers	(2) After Taxes and Transfers
Lowest 20 percent	1.9	8.0
Second 20 percent	7.6	11.3
Third 20 percent	13.6	15.0
Fourth 20 percent	21.8	20.6
Highest 20 percent	55.1	45.1

Source: Congressional Budget Office, **www.cbo.gov.** Income received "before taxes and transfers" excludes government cash transfers, realized capital gains, and employer-provided health insurance. Income received "after taxes and transfers" includes both cash and noncash transfers as well as realized capital gains and employer-provided health insurance. Numbers may not add to 100 percent due to rounding.

if all people receive exactly the same stream of income over their lifetimes, considerable income inequality would still exist in any specific year because of age differences. In any single year, the young and the old would receive low incomes while the middle-aged receive high incomes.

If we change from a "snapshot" view of income distribution in a single year to a "time exposure" portraying incomes over much longer periods, we find considerable movement of income receivers among income classes. For instance, one study showed that between 1996 and 2005, half the individuals in the lowest quintile of the U.S. income distribution in 1996 were in a higher income quintile in 2005. Almost 25 percent made it to the middle fifth and 5 percent achieved the top quintile. There was income mobility in both directions. About 57 percent of the top 1 percent of income receivers in 1996 had dropped out of that category by 2005. Overall, income mobility between 1996 and 2005 was the same as it was the previous 10 years. All this correctly suggests that income is more equally distributed over a 5-, 10-, or 20-year period than in any single year.[3]

In short, individual and family income mobility over time is significant; for many people, "low income" and "high income" are not permanent conditions. Also, the longer the time period considered, the more equal the distribution of income becomes.

[3]U.S. Department of the Treasury, *Income Mobility in the U.S. from 1996–2005*, November 13, 2007, pp. 1–22.

Effect of Government Redistribution

The income data in Table 21.1 and Figure 21.1 include wages, salaries, dividends, and interest. They also include all cash transfer payments such as Social Security, unemployment compensation benefits, and welfare assistance to needy families. The data are before-tax data and therefore do not take into account the effects of personal income and payroll (Social Security) taxes that are levied directly on income receivers. Nor do they include in-kind or **noncash transfers,** which provide specific goods or services rather than cash. Noncash transfers include such things as Medicare, Medicaid, housing subsidies, subsidized school lunches, and food stamps. Such transfers are "incomelike," since they enable recipients to "purchase" goods and services.

One economic function of government is to redistribute income, if society so desires. Figure 21.2 and its table[4] reveal that government significantly redistributes income from higher- to lower-income households through taxes and transfers. Note that the U.S. distribution of household income before taxes and transfers are taken into account (dark red Lorenz curve) is substantially less equal than the distribution after taxes and transfers (light red Lorenz curve). Without government redistribution,

[4]The "before" data in this table differ from the data in the table in Figure 21.1 because the latter include cash transfers. Also, the data in Figure 21.2 are based on a broader concept of income than are the data in Figure 21.1.

the lowest 20 percent of households in 2009 would have received only 1.9 percent of total income. *With* redistribution, they received 8.0 percent, or 4.2 times as much.

Which contributes more to redistribution, government taxes or government transfers? The answer is transfers. As discussed in Chapter 18's Last Word, the combined federal, state, and local tax system in the United States is only modestly progressive. As a result, nearly all the reduction in income inequality in the United States is attributable to transfer payments. Together with job opportunities, transfer payments have been the most important means of alleviating poverty in the United States.

Causes of Income Inequality

LO21.2 Discuss the extent and sources of income inequality.

There are several causes of income inequality in the United States. In general, the market system is permissive of a high degree of income inequality because it rewards individuals based on the contributions that they make, or the resources that they own, in producing society's output.

More specifically, the factors that contribute to income inequality are the following.

Ability

People have different mental, physical, and aesthetic talents. Some have inherited the exceptional mental qualities that are essential to such high-paying occupations as medicine, corporate leadership, and law. Others are blessed with the physical capacity and coordination to become highly paid professional athletes. A few have the talent to become great artists or musicians or have the beauty to become top fashion models. Others have very weak mental endowments and may work in low-paying occupations or may be incapable of earning any income at all. The intelligence and skills of most people fall somewhere in between.

Education and Training

Native ability alone rarely produces high income; people must develop and refine their capabilities through education and training. Individuals differ significantly in the amount of education and training they obtain and thus in their capacity to earn income. Such differences may be a matter of choice: Nguyen enters the labor force after graduating from high school, while Nyberg takes a job only after earning a college degree. Other differences may be involuntary: Nguyen and her parents may simply be unable to finance a college education.

People also receive varying degrees of on-the-job training, which contributes to income inequality. Some workers learn valuable new skills each year on the job and therefore experience significant income growth over time; others receive little or no on-the-job training and earn no more at age 50 than they did at age 30. Moreover, firms tend to select for advanced on-the-job training the workers who have the most formal education. That added training magnifies the education-based income differences between less-educated and better-educated individuals.

Discrimination

Discrimination in education, hiring, training, and promotion undoubtedly causes some income inequality. If discrimination confines certain racial, ethnic, or gender groups to lower-pay occupations, the supply of labor in those occupations will increase relative to demand, and hourly wages and income in those lower-pay jobs will decline. Conversely, labor supply will be artificially reduced in the higher-pay occupations populated by "preferred" workers, raising their wage rates and income. In this way, discrimination can add to income inequality. In fact, economists cannot account for all racial, ethnic, and gender differences in work earnings on the basis of differences in years of education, quality of education, occupations, and annual hours of work. Many economists attribute the unexplained residual to discrimination.

Economists, however, do not see discrimination by race, gender, and ethnicity as a dominant factor explaining income inequality. The income distributions *within* racial or ethnic groups that historically have been targets of discrimination—for example, African Americans—are similar to the income distribution for whites. Other factors besides discrimination are obviously at work.

Nevertheless, discrimination is an important concern since it harms individuals and reduces society's overall output and income. We will discuss it in more detail later in this chapter.

Preferences and Risks

Incomes also differ because of differences in preferences for market work relative to leisure, market work relative to work in the household, and types of occupations. People who choose to stay home with children, work part-time, or retire early usually have less income than those who make the opposite choices. Those who are willing to take arduous, unpleasant jobs (for example, underground mining or heavy construction), to work long hours with great intensity, or to "moonlight" will tend to earn more.

Individuals also differ in their willingness to assume risk. We refer here not only to the race-car driver or the professional boxer but also to the entrepreneur. Although many entrepreneurs fail, many of those who develop successful new products or services realize very substantial incomes. That contributes to income inequality.

Unequal Distribution of Wealth

Income is a *flow;* it represents a stream of wage and salary earnings, along with rent, interest, and profits, as depicted in Chapter 2's circular flow diagram. In contrast, wealth is a *stock*, reflecting at a particular moment the financial and real assets an individual has accumulated over time. A retired person may have very little income and yet own a home, mutual fund shares, and a pension plan that add up to considerable wealth. A new college graduate may be earning a substantial income as an accountant, middle manager, or engineer but has yet to accumulate significant wealth.

As you will discover in this chapter's Last Word, the ownership of wealth in the United States is more unequal than the distribution of income. This inequality of wealth leads to inequality in rent, interest, and dividends, which in turn contributes to income inequality. Those who own more machinery, real estate, farmland, stocks, and bonds and who have more money in savings accounts obviously receive greater income from that ownership than people with less or no such wealth.

Market Power

The ability to "rig the market" on one's own behalf also contributes to income inequality. For example, in *resource* markets certain unions and professional groups have adopted policies that limit the supply of their services, thereby boosting the incomes of those "on the inside." Also, legislation that requires occupational licensing for, say, doctors, dentists, and lawyers can bestow market power that favors the licensed groups. In *product* markets, "rigging the market" means gaining or enhancing monopoly power, which results in greater profit and thus greater income to the firms' owners.

Luck, Connections, and Misfortune

Other forces also play a role in producing income inequality. Luck and "being in the right place at the right time" have helped individuals stumble into fortunes. Discovering oil on a ranch, owning land along a proposed freeway interchange, and hiring the right press agent have accounted for some high incomes. Personal contacts and political connections are other potential routes to attaining high income.

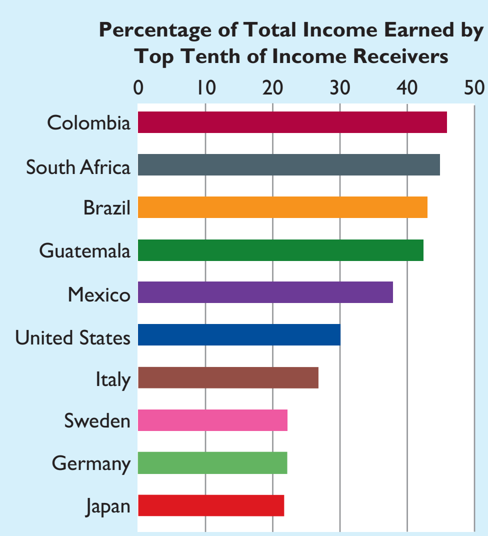
In contrast, economic misfortunes such as prolonged illness, serious accident, the death of the family breadwinner, or unemployment may plunge a family into the low range of income. The burden of such misfortune is borne very unevenly by the population and thus contributes to income inequality.

Income inequality of the magnitude we have described is not exclusively an American phenomenon. Global Perspective 21.1 compares income inequality (here by individuals, not by households) in the United States with that in several other nations. Income inequality tends to be greatest in South American nations, where land and capital resources are highly concentrated in the hands of a relatively small number of wealthy families.

GLOBAL PERSPECTIVE 21.1

Percentage of Total Income Received by the Top One-Tenth of Income Receivers, Selected Nations

The share of income going to the highest 10 percent of income receivers varies among nations.



Source: Human Development Report, 2009, **hdr.undp.org**. United Nations Development Programme. Used with permission of Palgrave Macmillan.

QUICK REVIEW 21.1

- Data reveal considerable income inequality in the United States; in 2011 the richest fifth of all households received 51.1 percent of before-tax income, and the poorest fifth received 3.2 percent.
- The Lorenz curve depicts income inequality graphically by comparing percentages of total families and percentages of total income. The Gini ratio is a measure of the overall dispersion of income and is found by dividing the area between the diagonal and the Lorenz curve by the total area below the diagonal.
- The distribution of income is less unequal over longer time periods.
- Government taxes and transfers significantly reduce income inequality by redistributing income from higher-income groups to lower-income groups; the bulk of this redistribution results from transfer payments.
- Differences in ability, education and training, preferences for market work versus nonmarket activities, property ownership, and market power—along with discrimination and luck—help explain income inequality.

Income Inequality over Time

LO21.3 Demonstrate how income inequality has changed since 1975.

Over a period of years economic growth has raised incomes in the United States: In *absolute* dollar amounts, the entire distribution of income has been moving upward. But incomes may move up in *absolute* terms while leaving the *relative* distribution of income less equal, more equal, or unchanged. Table 21.2 shows how the distribution of household income has changed since 1975. This income is "before tax" and includes cash transfers but not noncash transfers.

Rising Income Inequality since 1975

It is clear from Table 21.2 that the distribution of income by quintiles has become more unequal since 1975. In 2011 the lowest 20 percent of households received 3.8 percent of total before-tax income, compared with 4.4 in 1975. Meanwhile, the income share received by the highest 20 percent rose from 43.2 in 1975 to 48.9 percent in 2011. The percentage of income received by the top 5 percent of households also rose significantly over the 1975–2011 period.

Causes of Growing Inequality

Economists suggest several major explanations for the increase in U.S. income inequality since 1975.

Greater Demand for Highly Skilled Workers Perhaps the most significant contributor to the growing income inequality has been an increasing demand by many firms for workers who are highly skilled and well-educated. Moreover, several industries requiring highly skilled workers have either recently emerged or expanded greatly, such as the computer software, business consulting, biotechnology, health care, and Internet industries. Because highly skilled workers remain relatively scarce, their wages have been bid up. Consequently, the wage differences between them and less-skilled workers have increased.

Between 1980 and 2007 the wage difference between college graduates and high school graduates rose from 28 percent to 49 percent for women and from 22 percent to 44 percent for men. And the so-called *90-10 ratio*—how many times larger the hourly wage at the 90th percentile is compared to the hourly wage at the 10th percentile—rose from 3.6 in 1980 to 4.5 in 2007.[5]

[5]Economic Policy Institute, **www.epinet.org**. The college wage premiums are adjusted for differences in earnings based on race, ethnicity, marital status, and region.

TABLE 21.2 **Percentage of Total Before-Tax Income Received by Each One-Fifth, and by the Top 5 Percent, of Households, Selected Years**

Quintile	1975	1980	1985	1990	1995	2000	2005	2011
Lowest 20 percent	4.4	4.3	4.0	3.9	3.7	3.6	3.4	3.8
Second 20 percent	10.5	10.3	9.7	9.6	9.1	8.9	8.6	9.3
Third 20 percent	17.1	16.9	16.3	15.9	15.2	14.8	14.6	15.1
Fourth 20 percent	24.8	24.9	24.6	24.0	23.3	23.0	23.0	23.0
Highest 20 percent	43.2	43.7	45.3	46.6	48.7	49.8	50.4	48.9
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Top 5 percent	15.9	15.8	17.0	18.6	21.0	22.1	22.2	21.3

Source: Bureau of the Census, **www.census.gov**. Numbers may not add to 100 percent due to rounding.

The rising demand for skill has also shown up in rapidly rising pay for chief executive officers (CEOs), sizable increases in income from stock options, substantial increases in income for professional athletes and entertainers, and huge fortunes for successful entrepreneurs. This growth of "superstar" pay has also contributed to rising income inequality.

Demographic Changes The entrance of large numbers of less-experienced and less-skilled "baby boomers" into the labor force during the 1970s and 1980s may have contributed to greater income inequality in those two decades. Because younger workers tend to earn less income than older workers, their growing numbers contributed to income inequality. There has also been a growing tendency for men and women with high earnings potential to marry each other, thus increasing household income among the highest income quintiles. Finally, the number of households headed by single or divorced women has increased greatly. That trend has increased income inequality because such households lack a second major wage earner and also because the poverty rate for female-headed households is very high.

International Trade, Immigration, and Decline in Unionism Other factors are probably at work as well. Stronger international competition from imports has reduced the demand for and employment of less skilled (but highly paid) workers in such industries as the automobile and steel industries. The decline in such jobs has reduced the average wage for less-skilled workers. It also has swelled the ranks of workers in already low-paying industries, placing further downward pressure on wages there.

Similarly, the transfer of jobs to lower-wage workers in developing countries has exerted downward wage pressure on less-skilled workers in the United States. Also, an upsurge in the immigration of unskilled workers has increased the number of low-income households in the United States. Finally, the decline in unionism in the United States has undoubtedly contributed to wage inequality, since unions tend to equalize pay within firms and industries.

Two cautions: First, when we note growing income inequality, we are not saying that the "rich are getting richer and the poor are getting poorer" in terms of absolute income. Both the rich and the poor are experiencing rises in real income. Rather, what has happened is that, while incomes have risen in all quintiles, income growth has been fastest in the top quintile. Second, increased income inequality is not solely a U.S. phenomenon. The recent rise of inequality has also occurred in several other industrially advanced nations.

CONSIDER THIS . . .

Laughing at Shrek

Some economists say that the distribution of annual *consumption* is more meaningful for examining inequality of well-being than is the distribution of annual *income*. In a given year, people's consumption of goods and services may be above or below their income because they can save, draw down past savings, use credit cards, take out home mortgages, spend from inheritances, give money to charities, and so on.

A recent study of the distribution of consumption finds that annual consumption inequality is less than income inequality. Moreover, consumption inequality has remained relatively constant over several decades, even though income inequality has increased.*

The Economist magazine extends the argument even further, pointing out that despite the recent increase in income inequality, the products consumed by the rich and the poor are far closer in functionality today than at any other time in history.

> More than 70 percent of Americans under the official poverty line own at least one car. And the distance between driving a used Hyundai Elantra and new Jaguar XJ is well nigh undetectable compared with the difference between motoring and hiking through the muck. . . . A wide screen plasma television is lovely, but you do not need one to laugh at "Shrek". . . .
>
> Those intrepid souls who make vast fortunes turning out ever higher-quality goods at ever lower prices widen the income gap while reducing the differences that really matter.†

Economists generally agree that products and experiences once reserved exclusively for the rich in the United States have, in fact, become more commonplace for nearly all income classes. But skeptics argue that *The Economist*'s argument is too simplistic. Even though both are water outings, there is a fundamental difference between yachting among the Greek isles on your private yacht and paddling on a local pond in your kayak.

*Dirk Krueger and Fabrizio Perri, "Does Income Inequality Lead to Consumption Inequality?" *Review of Economic Studies,* 2006, pp. 163–193.

†*The Economist,* "Economic Focus: The New (Improved) Gilded Age," December 22, 2007, p. 122. © The Economist Newspaper Limited, London (2007).

The Lorenz curve can be used to contrast the distribution of income at different points in time. If we plotted Table 21.2's data as Lorenz curves, we would find that the curves shifted farther away from the diagonal between 1975 and 2011. The Gini ratio rose from 0.397 in 1975 to 0.477 in 2011.

Equality versus Efficiency

LO21.4 Debate the economic arguments for and against income inequality.

The main policy issue concerning income inequality is how much is necessary and justified. While there is no general agreement on the justifiable amount, we can gain insight by exploring the cases for and against greater equality.

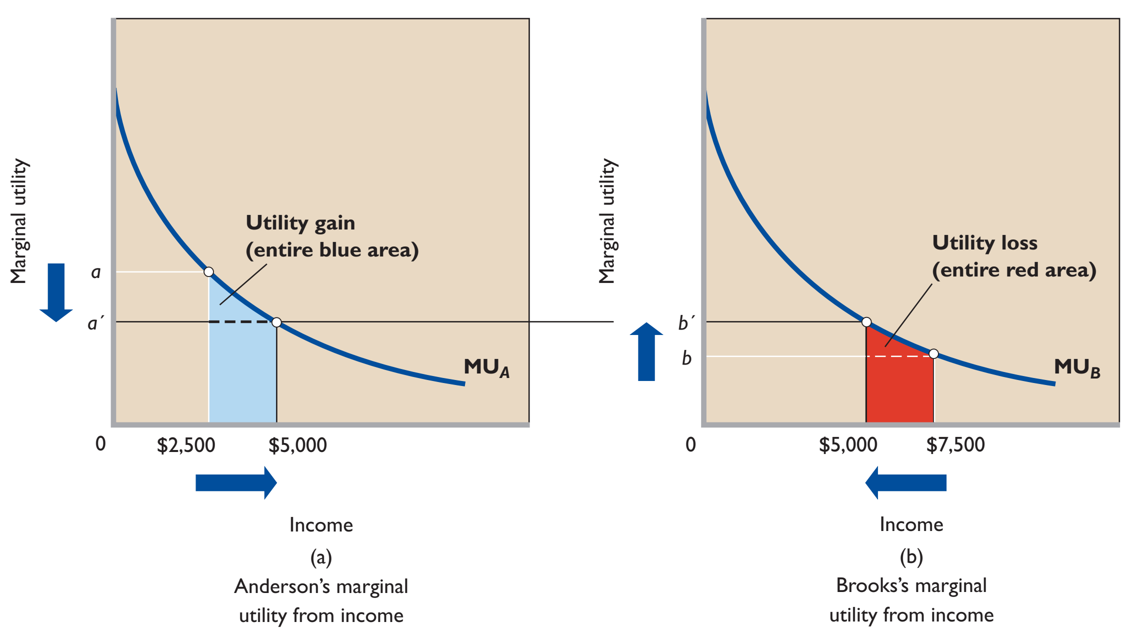
The Case for Equality: Maximizing Total Utility

The basic argument for an equal distribution of income is that income equality maximizes total consumer satisfaction (utility) from any particular level of output and income. The rationale for this argument is shown in Figure 21.3, in which we assume that the money incomes of two individuals, Anderson and Brooks, are subject to diminishing marginal utility. In any time period, income receivers spend the first dollars received on the products they value most—products whose marginal utility is high. As their most pressing wants become satisfied, consumers then spend additional dollars of income on less important, lower-marginal-utility goods. The identical diminishing-marginal-utility-from-income curves (MU_A and MU_B in the figure) reflect the assumption that Anderson and Brooks have the same capacity to derive utility from income.

Now suppose that there is \$10,000 worth of income (output) to be distributed between Anderson and Brooks. According to proponents of income equality, the optimal distribution is an equal distribution, which causes the marginal utility of the last dollar spent to be the same for both persons. We can prove this by demonstrating that if the income distribution is initially unequal, then distributing income more equally can increase the combined utility of the two individuals.

Suppose that the \$10,000 of income initially is distributed unequally, with Anderson getting \$2,500 and Brooks \$7,500. The marginal utility, *a*, from the last dollar received by Anderson is high, and the marginal utility, *b*, from Brooks's last dollar of income is low. If a single dollar of income is shifted from Brooks to Anderson—that is,

FIGURE 21.3 The utility-maximizing distribution of income. With identical marginal-utility-of-income curves MU_A and MU_B, Anderson and Brooks will maximize their combined utility when any amount of income (say, \$10,000) is equally distributed. If income is unequally distributed (say, \$2,500 to Anderson and \$7,500 to Brooks), the marginal utility derived from the last dollar will be greater for Anderson than for Brooks, and a redistribution of income toward equality will result in a net increase in total utility. The utility gained by equalizing income at \$5,000 each, shown by the blue area below curve MU_A in panel (a), exceeds the utility lost, indicated by the red area below curve MU_B in (b).



toward greater equality—then Anderson's utility increases by *a* and Brooks's utility decreases by *b*. The combined utility then increases by *a* minus *b* (Anderson's large gain minus Brooks's small loss). The transfer of another dollar from Brooks to Anderson again increases their combined utility, this time by a slightly smaller amount. Continued transfer of dollars from Brooks to Anderson increases their combined utility until the income is evenly distributed and both receive $5,000. At that time their marginal utilities from the last dollar of income are equal (at a' and b'), and any further income redistribution beyond the $2,500 already transferred would begin to create inequality and decrease their combined utility.

The area under the MU curve and to the left of the individual's particular level of income represents the total utility of that income. Therefore, as a result of the transfer of the $2,500, Anderson has gained utility represented by the blue area below curve MU_A, and Brooks has lost utility represented by the red area below curve MU_B. The blue area is obviously greater than the red area, so income equality yields greater combined total utility than income inequality does.

The Case for Inequality: Incentives and Efficiency

Although the logic of the argument for equality might seem solid, critics of income equality say that the transfer of income required to create income equality is both unfair and unwise. They say that income inequality largely reflects rewards to individuals for supplying their talents and resources to the economy. They conclude that it is not fair to take some of Brooks's income and give it to Anderson. Further, critics of income equality point out that proponents of income equality falsely assume that there is some fixed amount of output produced and therefore income to be distributed. These critics argue that the way in which income is distributed is an important determinant of the amount of output or income that is produced and is available for distribution.

Suppose once again in Figure 21.3 that Anderson earns $2,500 and Brooks earns $7,500. In moving toward equality, society (the government) must tax away some of Brooks's income and transfer it to Anderson. This tax and transfer process diminishes the income rewards of high-income Brooks and raises the income rewards of low-income Anderson; in so doing, it reduces the incentives of both to earn high incomes. Why should high-income Brooks work hard, save and invest, or undertake entrepreneurial risks when the rewards from such activities will be reduced by taxation? And why should low-income Anderson be motivated to increase his income through market activities when the government stands ready to transfer income to him? Taxes are a reduction in the rewards from increased productive effort; redistribution through transfers is a reward for diminished effort.

In the extreme, imagine a situation in which the government levies a 100 percent tax on income and distributes the tax revenue equally to its citizenry. Why would anyone work hard? Why would anyone work at all? Why would anyone assume business risk? Or why would anyone save (forgo current consumption) in order to invest? The economic incentives to "get ahead" will have been removed, greatly reducing society's total production and income. That is, the way income is distributed affects the size of that income. The basic argument for income inequality therefore is twofold: (1) income inequality is justified because it results from differences in the quantity and quality of labor and other resources supplied by individuals to the

CONSIDER THIS . . .

Slicing the Pizza

The equality-efficiency tradeoff might better be understood through an analogy. Assume that society's income is a huge pizza, baked year after year, *with the sizes of the pieces going to people on the basis of their contribution to making it.* Now suppose that for fairness reasons, society decides some people are getting pieces that are too large and others are getting pieces too small. But when society redistributes the pizza to make the sizes more equal, they discover the result is a smaller pizza than before. Why participate in making the pizza if you get a decent-size piece without contributing? The shrinkage of the pizza represents the efficiency loss—the loss of output and income—caused by the harmful effects of the redistribution on incentives to work, to save and invest, and to accept entrepreneurial risk. The shrinkage also reflects the resources that society must divert to the bureaucracies that administer the redistribution system.

How much pizza shrinkage will society accept while continuing to agree to the redistribution? If redistributing pizza to make it less unequal reduces the size of the pizza, what amount of pizza loss will society tolerate? Is a loss of 10 percent acceptable? 25 percent? 75 percent? This is the basic question in any debate over the ideal size of a nation's income redistribution program.

economy and (2) it is an unavoidable consequence of maintaining the incentives needed to motivate people to produce output and income year after year.

The Equality-Efficiency Trade-off

At the essence of the income equality-inequality debate is a fundamental trade-off between equality and efficiency. In this **equality-efficiency trade-off,** greater income equality (achieved through redistribution of income) comes at the opportunity cost of reduced production and income. And greater production and income (through reduced redistribution) comes at the expense of less equality of income. The trade-off obligates society to choose how much redistribution it wants, in view of the costs. If society decides it wants to redistribute income, it needs to determine methods that minimize the adverse effects on fairness, incentives, productivity, and economic efficiency.

The Economics of Poverty

LO21.5 Relate how poverty is measured and its incidence by age, gender, ethnicity, and other characteristics.

We now turn from the broader issue of income distribution to the more specific issue of very low income, or "poverty." A society with a high degree of income inequality can have a high, moderate, or low amount of poverty. We therefore need a separate examination of poverty.

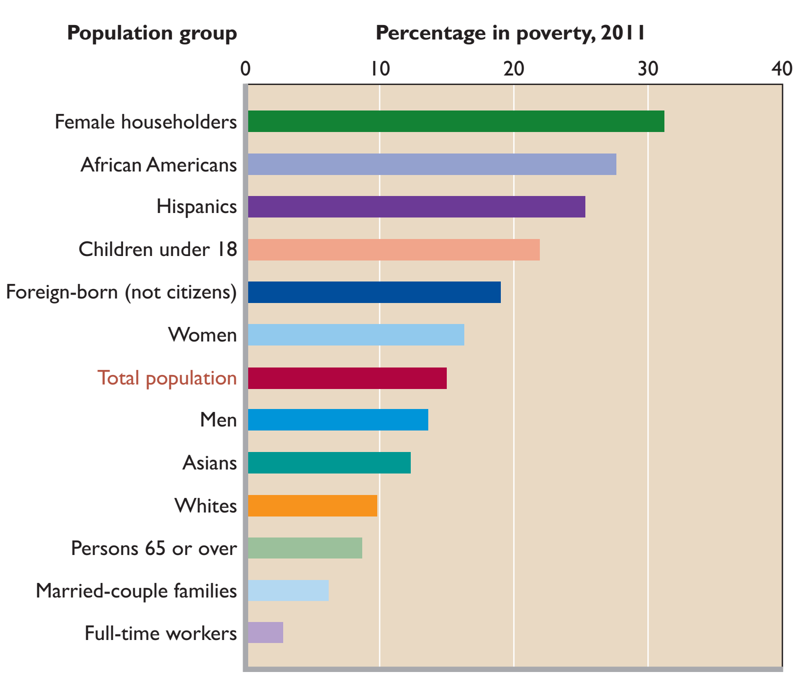
Definition of Poverty

Poverty is a condition in which a person or a family does not have the means to satisfy basic needs for food, clothing, shelter, and transportation. The means include currently earned income, transfer payments, past savings, and property owned. The basic needs have many determinants, including family size and the health and age of its members.

The federal government has established minimum income thresholds below which a person or a family is "in poverty." In 2011 an unattached individual receiving less than $11,702 per year was said to be living in poverty. For a family of four, the poverty line was $22,891; for a family of six, it was $29,494. Based on these thresholds, in 2011 about 46.2 million Americans lived in poverty. In 2011 the **poverty rate**—the percentage of the population living in poverty—was 15 percent.

Incidence of Poverty

The poor are heterogeneous: They can be found in all parts of the nation; they are of all races and ethnicities, rural and urban, young and old. But as Figure 21.4 indicates, poverty is far from randomly distributed. For example, the poverty rate for African Americans is above the national average, as is the rate for Hispanics, while the rates for whites and Asians are below the average. In 2011, the poverty rates for African Americans and Hispanics



Source: Bureau of the Census, **www.census.gov**.

FIGURE 21.4 Poverty rates among selected population groups, 2011. Poverty is disproportionately borne by African Americans, Hispanics, children, foreign-born residents who are not citizens, and families headed by women. People who are employed full-time, have a college degree, or are married tend to have low poverty rates.

were 27.6 and 25.3, respectively; for whites and Asians, 9.8 and 12.3 percent, respectively.

Figure 21.4 shows that female-headed households (no husband present), foreign-born noncitizens, and children under 18 years of age have very high incidences of poverty. Marriage and full-time, year-round work are associated with low poverty rates, and, because of the Social Security system, the incidence of poverty among the elderly is less than that for the population as a whole.

The high poverty rate for children is especially disturbing because poverty tends to breed poverty. Poor children are at greater risk for a range of long-term problems, including poor health and inadequate education, crime, drug use, and teenage pregnancy. Many of today's impoverished children will reach adulthood unhealthy and illiterate and unable to earn above-poverty incomes.

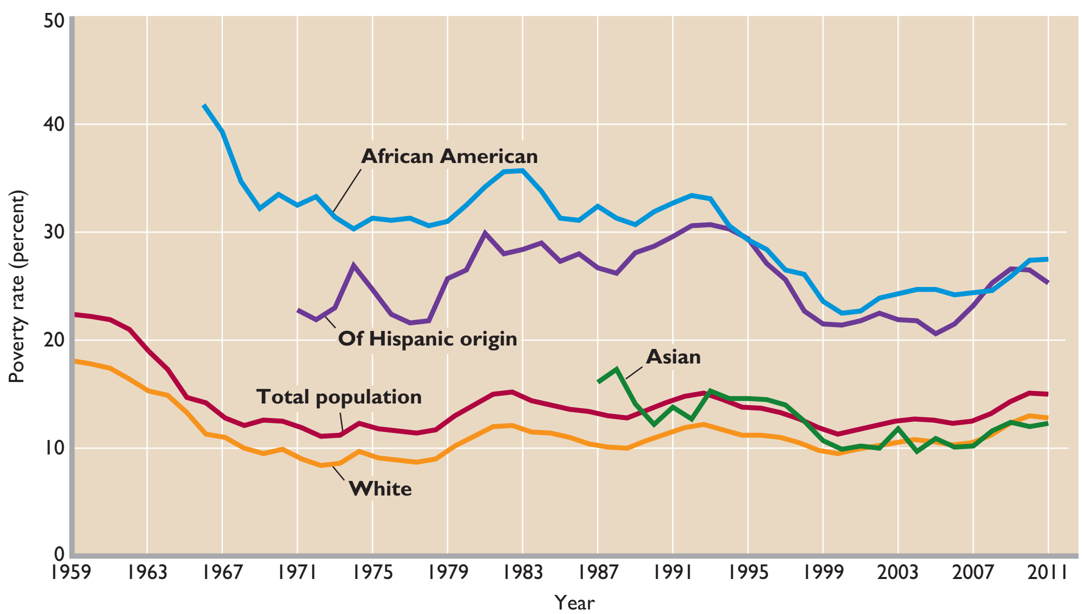
As many as half of people in poverty are poor for only 1 or 2 years before climbing out of poverty. But poverty is much more long-lasting among some groups than among others. In particular, African-American and Hispanic families, families headed by women, persons with little education and few labor market skills, and people who are dysfunctional because of drug use, alcoholism, or mental illness are more likely than others to remain in poverty. Also, long-lasting poverty is heavily present in depressed areas of cities, parts of the Deep South, and some Indian reservations.

Poverty Trends

As Figure 21.5 shows, the total poverty rate fell significantly between 1959 and 1969, stabilized at 11 to 13 percent over the next decade, and then rose in the early 1980s. In 1993 the rate was 15.1 percent, the highest since 1983. Between 1993 and 2000 the rate turned downward, falling to 11.3 percent in 2000. Because of recession, slow employment growth, and relatively slow wage growth, the poverty rate rose from 11.7 percent in 2001 to 12.7 percent in 2004. During the second half of the 1990s, poverty rates plunged for African Americans, Hispanics, and Asians, and they have remained historically low. Nevertheless, in 2008 African Americans and Hispanics still had poverty rates that were roughly double the rate for whites.

The recession that began in December 2007 increased poverty rates for all groups with, for instance, the Asian poverty rate rising from 10.2 percent in 2007 to 11.8 percent in 2008. As data became available for the years 2009 to 2011, many economists were surprised that poverty rates appeared

FIGURE 21.5 Poverty-rate trends, 1959–2011. Although the national poverty rate declined sharply between 1959 and 1969, it stabilized in the 1970s only to increase significantly in the early 1980s. Between 1993 and 2000 it substantially declined, before rising slightly again in the immediate years following the 2001 recession. Although poverty rates for African Americans and Hispanics are much higher than the average, they significantly declined during the 1990s. Poverty rates rose in 2008 in response to the recession that began in December 2007.



Source: Bureau of the Census, **www.census.gov**.

to level off or fall despite the widespread and lingering unemployment caused by the so-called Great Recession.

Measurement Issues

The poverty rates and trends in Figures 21.4 and 21.5 should be interpreted cautiously. The official income thresholds for defining poverty are necessarily arbitrary and therefore may inadequately measure the true extent of poverty in the United States.

Some observers say that the high cost of living in major metropolitan areas means that the official poverty thresholds exclude millions of families whose income is slightly above the poverty level but clearly inadequate to meet basic needs for food, housing, and medical care. These observers use city-by-city studies on "minimal income needs" to argue that poverty in the United States is much more widespread than officially measured and reported.

In contrast, some economists point out that using income to measure poverty understates the standard of living of many of the people who are officially poor. When individual, household, or family *consumption* is considered rather than family *income*, some of the poverty in the United States disappears. Some low-income families maintain their consumption by drawing down past savings, borrowing against future income, or selling homes. Moreover, many poverty families receive substantial noncash benefits such as food stamps and rent subsidies that boost their living standards. Such "in-kind" benefits are not included in determining a family's official poverty status.

The U.S. Income-Maintenance System

LO21.6 Identify the major components of the income-maintenance program in the United States.

Regardless of how poverty is measured, economists agree that considerable poverty exists in the United States. Helping those who have very low income is a widely accepted goal of public policy. A wide array of antipoverty programs, including education and training programs, subsidized employment, minimum-wage laws, and antidiscrimination policies, are designed to increase the earnings of the poor. In addition, a number of income-maintenance programs were devised to reduce poverty; the most important are listed in Table 21.3. These programs involve large expenditures and have numerous beneficiaries.

TABLE 21.3 **Characteristics of Major Income-Maintenance Programs**

Program	Basis of Eligibility	Source of Funds	Form of Aid	Expenditures,* Billions	Beneficiaries, Millions
Social Insurance Programs					
Social Security	Age, disability, or death of parent or spouse; life-time work earnings	Federal payroll tax on employers and employees	Cash	$736	51
Medicare	Age or disability	Federal payroll tax on employers and employees	Subsidized health insurance	$523	63
Unemployment compensation	Unemployment	State and federal payroll taxes on employers	Cash	$55	14
Public Assistance Programs					
Supplemental Security Income (SSI)	Age or disability; income	Federal revenues	Cash	$48	8
Temporary Assistance for Needy Families (TANF)	Certain families with children; income	Federal-state-local revenues	Cash and services	$15	5
Supplemental Nutrition Assistance Program (SNAP)	Income	Federal revenues	Cash via EBT cards	$78	45
Medicaid	Persons eligible for TANF or SSI and medically indigent	Federal-state-local revenues	Subsidized medical services	$404	54
Earned-income tax credit (EITC)	Low-wage working families	Federal revenues	Refundable tax credit, cash	$62	27

*Expenditures by federal, state, and local governments; excludes administrative expenses.

Source: Social Security Administration, *Annual Statistical Supplement, 2011,* **www.socialsecurity.gov**; U.S. Department of Agriculture, **www.fns.usda.gov**; Internal Revenue Service, **www.irs.gov/taxstats**; and other government sources. Latest data.

The U.S. income-maintenance system consists of two kinds of programs: (1) social insurance and (2) public assistance, or "welfare." Both are known as **entitlement programs** because all eligible persons are legally entitled to receive the benefits set forth in the programs.

Social Insurance Programs

Social insurance programs partially replace earnings that have been lost due to retirement, disability, or temporary unemployment; they also provide health insurance for the elderly. The main social insurance programs are Social Security, unemployment compensation, and Medicare. Benefits are viewed as earned rights and do not carry the stigma of public charity. These programs are financed primarily out of federal payroll taxes. In these programs the entire population shares the risk of an individual's losing income because of retirement, unemployment, disability, or illness. Workers (and employers) pay a part of wages into a government fund while they are working. The workers are then entitled to benefits when they retire or when a specified misfortune occurs.

Social Security and Medicare The major social insurance program is known as **Social Security.** It is a federal pension program that replaces part of the earnings lost when workers retire, become disabled, or die. This gigantic program ($736 billion in 2011) is financed by compulsory payroll taxes levied on both employers and employees. Workers currently may retire at age 65 and receive full retirement benefits or retire early at age 62 with reduced benefits. When a worker dies, benefits accrue to his or her family survivors. Special provisions provide benefits for disabled workers.

Social Security covers over 90 percent of the workforce; some 51 million people receive Social Security benefits, with benefits for retirees averaging about $1,230 per month. In 2011, those benefits were financed with a combined Social Security and Medicare payroll tax of 15.3 percent, with both the worker and the employer paying 7.65 percent on their first $110,100 of earnings. The 7.65 percent tax comprises 6.2 percent for Social Security and 1.45 percent for Medicare. Self-employed workers pay a tax of 15.3 percent.

Medicare is a federal insurance program that provides health insurance benefits to those 65 or older and people who are disabled. It is financed by payroll taxes on employers and employees. This overall 2.9 percent tax is paid on all work income, not just on the first $110,100. Medicare also makes available supplementary low-cost insurance programs that help pay for doctor visits and prescription drug expenses. In 2011, some 63 million people received Medicare benefits. The benefits paid to recipients totaled $523 billion.

The number of retirees drawing Social Security and Medicare benefits is rapidly rising relative to the number of workers paying payroll taxes. As a result, Social Security and Medicare face serious long-term funding problems. These fiscal imbalances have spawned calls to reform the programs.

Unemployment Compensation All 50 states sponsor unemployment insurance programs called **unemployment compensation,** a federal–state social insurance program that makes income available to workers who are unemployed. This insurance is financed by a relatively small payroll tax, paid by employers, which varies by state and by the size of the firm's payroll. Any insured worker who becomes unemployed can, after a short waiting period, become eligible for benefit payments. The program covers almost all wage and salary workers. The size of payments and the number of weeks they are made available vary considerably from state to state. Generally, benefits approximate 33 percent of a worker's wages up to a certain maximum payment. In 2011 benefits averaged about $297 weekly. The number of beneficiaries and the level of total disbursements vary with economic conditions.

Typically, unemployment compensation payments last for a maximum of 26 weeks. But during recessions—when unemployment rates soar—Congress extends the benefits for additional weeks.

Public Assistance Programs

Public assistance programs (welfare) provide benefits to people who are unable to earn income because of permanent disabling conditions or who have no or very low income and also have dependent children. These programs are financed out of general tax revenues and are regarded as public charity. They include "means tests" that require that individuals and families demonstrate low incomes in order to qualify for aid. The federal government finances about two-thirds of the welfare program expenditures, and the rest is paid for by the states.

Many needy persons who do not qualify for social insurance programs are assisted through the federal government's **Supplemental Security Income (SSI)** program. This is a federal program (financed by general tax revenues) that provides a uniform nationwide minimum income for the aged, blind, and disabled who are unable to work and who do not qualify for Social Security aid. In 2013 the average monthly payment was $710 for individuals and $1,066 for couples with both people eligible. More than half the states provide additional income supplements to the aged, blind, and disabled.

Temporary Assistance for Needy Families (TANF) is the basic welfare program for low-income families in the United States. The program is financed through general federal tax revenues and consists of lump-sum payments of federal money to states to operate their own welfare and work programs. These lump-sum payments are called TANF funds, and in 2011 about 4.4 million people (including children) received TANF assistance. TANF expenditures in 2011 were about $15 billion.

In 1996 TANF replaced the six-decade-old Aid for Families with Dependent Children (AFDC) program. Unlike that welfare program, TANF established work requirements and placed limits on the length of time a family can receive welfare payments. Specifically, the TANF program:

- Set a lifetime limit of 5 years on receiving TANF benefits and requires able-bodied adults to work after receiving assistance for 2 years.
- Ended food-stamp eligibility for able-bodied persons age 18 to 50 (with no dependent children) who are not working or engaged in job-training programs.
- Tightened the definition of "disabled children" as it applies for eligibility of low-income families for Supplemental Security Income (SSI) assistance.
- Established a 5-year waiting period on public assistance for new legal immigrants who have not become citizens.

In 1996, about 12.6 million people, or about 4.8 percent of the U.S. population, were welfare recipients. By the middle of 2007 those totals had declined to 3.9 million and 1.3 percent of the population. The recession that began in December 2007 pushed the number of welfare recipients up to about 4.4 million by December 2009. These recipients accounted for about 1.4 percent of the population in December 2009.

The welfare program has greatly increased the employment rate (= employment/population) for single mothers with children under age 6—a group particularly prone to welfare dependency. Today, that rate is about 13 percentage points higher than it was in 1996.

The **Supplemental Nutrition Assistance Program (SNAP)** was formerly known as the food-stamp program. SNAP is a federal program (financed through general tax revenues) that permits eligible low-income persons to obtain vouchers that can be used to buy food. It is designed to provide all low-income Americans with a nutritionally adequate diet. Under the program, eligible households receive monthly deposits of spendable electronic money on specialized debit cards known as Electronic Benefit Transfer (EBT) cards. The EBT cards are designed so that the deposits can be spent only on food. The amount deposited onto a family's EBT card varies inversely with the family's earned income.

Medicaid is a federal program (financed by general tax revenues) that provides medical benefits to people covered by the SSI and TANF (basic welfare) programs. It helps finance the medical expenses of individuals participating in those programs.

The **earned-income tax credit (EITC)** is a refundable federal tax credit provided to low-income wage earners to supplement their families' incomes and encourage work. It is available for low-income working families, with or without children. The credit reduces the federal income taxes that such families owe or provides them with cash payments if the credit exceeds their tax liabilities. The purpose of the credit is to offset Social Security taxes paid by low-wage earners and thus keep the federal government from "taxing families into poverty." In essence, EITC is a wage subsidy from the federal government that works out to be as much as $2 per hour for the lowest-paid workers with families. Under the program many people owe no income tax and receive direct checks from the federal government once a year. According to the Internal Revenue Service, 27 million taxpayers received $62 billion in payments from the EITC in 2011.

Several other welfare programs are not listed in Table 21.3. Some provide help in the form of noncash transfers. Head Start provides education, nutrition, and social services to economically disadvantaged 3- and 4-year-olds. Housing assistance in the form of rent subsidies and funds for construction is available to low-income families. Pell grants provide assistance to undergraduate students from low-income families. Low-income home energy assistance provides help with home heating bills. Other programs—such as veteran's assistance and black lung benefits—provide cash assistance to those eligible.

QUICK REVIEW 21.2

- The basic argument for income equality is that it maximizes total utility by equalizing the marginal utility of the last dollar of income received by all people.
- The basic argument for income inequality is that it is an unavoidable consequence of maintaining the economic incentives for production.
- By government standards, 46.2 million people in the United States, or 15 percent of the population, lived in poverty in 2011.
- The U.S. income-maintenance system includes both social insurance programs and public assistance (welfare) programs.

Economic Analysis of Discrimination

LO21.7 Discuss labor market discrimination and how it might affect hiring decisions and wages.

Although the majority of Americans who are in the lowest income quintile or in poverty are white, African Americans and Hispanics are in those two categories disproportionally to their total populations. For that reason, the percentages of all African Americans and Hispanics receiving public assistance from the TANF, SSI, and food stamp programs are also well above the average for the entire population. This fact raises the question of what role, if any, discrimination plays in reducing wages for some and increasing wages for others.

Discrimination is the practice of according people inferior treatment (for example, in hiring, occupational access, education and training, promotion, wage rate, or working conditions) on the basis of some factor such as race, gender, or ethnicity. People who practice discrimination are said to exhibit a prejudice or a bias against the groups they discriminate against.

Prejudice reflects complex, multifaceted, and deeply ingrained beliefs and attitudes. Thus, economics can contribute some insights into discrimination but no detailed explanations. With this caution in mind, let's look more deeply into the economics of discrimination.

Taste-for-Discrimination Model

The **taste-for-discrimination model** examines prejudice by using the emotion-free language of demand theory. It views discrimination as resulting from a preference or taste for which the discriminator is willing to pay. The model assumes that, for whatever reason, prejudiced people experience a subjective or psychic cost—a disutility—whenever they must interact with those they are biased against. Consequently, they are willing to pay a certain "price" to avoid interactions with the nonpreferred group. The size of this price depends directly on the degree of prejudice.

ORIGIN OF THE IDEA

O21.1
Taste-for-discrimination model

The taste-for-discrimination model is general, since it can be applied to race, gender, age, and religion. But our discussion focuses on employer discrimination, in which employers discriminate against nonpreferred workers. For concreteness, we will look at a white employer discriminating against African-American workers.

Discrimination Coefficient A prejudiced white employer behaves as if employing African-American workers would add a cost. The amount of this cost—this disutility—is reflected in a **discrimination coefficient,** *d*, measured in monetary units. Because the employer is not prejudiced against whites, the cost of employing a white worker is the white wage rate, W_w. However, the employer's perceived "cost" of employing an African-American worker is the African-American worker's wage rate, W_{aa}, *plus* the cost *d* involved in the employer's prejudice, or $W_{aa} + d$.

The prejudiced white employer will have no preference between African-American and white workers when the total cost per worker is the same, that is, when $W_w = W_{aa} + d$. Suppose the market wage rate for whites is \$10 and the monetary value of the disutility the employer attaches to hiring African Americans is \$2 (that is, $d = \$2$). This employer will be indifferent between hiring African Americans and whites only when the African-American wage rate is \$8, since at this wage the perceived cost of hiring either a white or an African-American worker is \$10:

$$\$10 \text{ white wage} = \$8 \text{ African-American wage} + \$2 \text{ discrimination coefficient}$$

It follows that our prejudiced white employer will hire African Americans only if their wage rate is sufficiently below that of whites. By "sufficiently" we mean at least the amount of the discrimination coefficient.

The greater a white employer's taste for discrimination as reflected in the value of *d*, the larger the difference between white wages and the lower wages at which African Americans will be hired. A "color-blind" employer whose *d* is \$0 will hire equally productive African Americans and whites impartially if their wages are the same. A blatantly prejudiced white employer whose *d* is infinity would refuse to hire African Americans even if the African-American wage were zero.

Most prejudiced white employers will not refuse to hire African Americans under all conditions. They will, in fact, *prefer* to hire African Americans if the actual white-black wage difference in the market exceeds the value of *d*. In our example, if whites can be hired at \$10 and equally productive African Americans at only \$7.50, the biased white employer will hire African Americans. That employer is willing to pay a wage difference of up to \$2 per hour for whites to satisfy his or her bias, but no more. At the \$2.50 actual difference, the employer will hire African Americans.

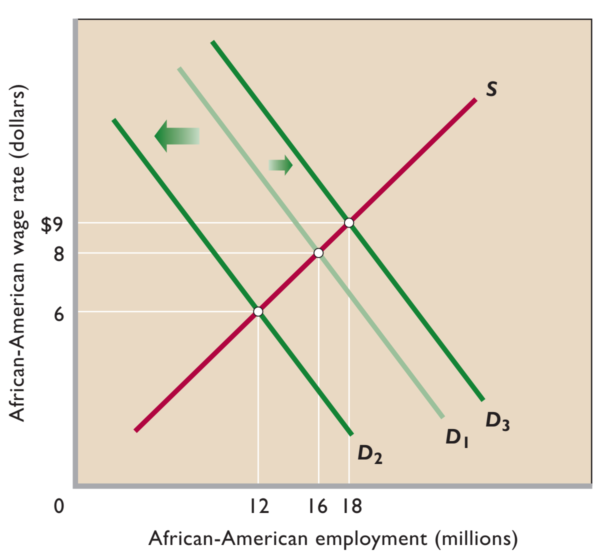
Conversely, if whites can be hired at \$10 and African Americans at \$8.50, whites will be hired. Again, the biased employer is willing to pay a wage difference of up to \$2 for whites; a \$1.50 actual difference means that hiring whites is a "bargain" for this employer.

Prejudice and the Market African-American–White Wage Ratio For a particular supply of African-American workers, the actual African-American–white wage ratio—the ratio determined in the labor market—will depend on the collective prejudice of white employers. To see why, consider Figure 21.6, which shows a labor market for *African-American* workers. Initially, suppose the relevant labor demand curve is D_1, so the equilibrium *African-American* wage is \$8 and the equilibrium level of African-American employment is 16 million. If we assume that the *white* wage (not shown) is \$10, then the initial African-American–white wage ratio is 0.8 (= \$8/\$10).

Now assume that prejudice against African-American workers increases—that is, the collective *d* of white employers rises. An increase in *d* means an increase in the perceived cost of African-American labor at each African-American wage rate, and that reduces the demand for African-American labor, say, from D_1 to D_2. The African-American wage rate falls from \$8 to \$6 in the market, and the level of African-American employment declines from 16 million to 12 million. The increase in white employer prejudice reduces the African-American wage rate and thus the actual African-American–white wage ratio. If the white wage rate remains at \$10, the new African-American–white ratio is 0.6 (= \$6/\$10).

Conversely, suppose social attitudes change such that white employers become less biased and their discrimination coefficient as a group declines. This decreases the perceived cost of African-American labor at each African-American wage rate, so the demand for African-American labor increases, as from D_1 to D_3. In this case, the African-American wage rate rises to \$9, and employment of African-American workers increases to 18 million. The decrease in white employer prejudice increases the African-American wage rate and thus the actual African-American–white wage ratio. If the white wage remains at \$10, the new African-American–white wage ratio is 0.9 (= \$9/\$10).

FIGURE 21.6 The African-American wage and employment level in the taste-for-discrimination model. An increase in prejudice by white employers as reflected in higher discrimination coefficients would decrease the demand for African-American workers, here from D_1 to D_2, and reduce the African-American wage rate and level of African-American employment. Not shown, this drop in the African-American wage rate would lower the African-American–white wage ratio. In contrast, if prejudice were reduced such that discrimination coefficients of employers declined, the demand for African-American labor would increase, as from D_1 to D_3, boosting the African-American wage rate and level of employment. The higher African-American wage rate would increase the African-American–white wage ratio.



Competition and Discrimination The taste-for-discrimination model suggests that competition will reduce discrimination in the very long run, as follows: The actual African-American–white wage difference for equally productive workers—say, \$2—allows nondiscriminators to hire African Americans for less than whites. Firms that hire African-American workers will therefore have lower actual wage costs per unit of output and lower average total costs than will the firms that discriminate. These lower costs will allow nondiscriminators to underprice discriminating competitors, eventually driving them out of the market.

But critics of this implication of the taste-for-discrimination model say that it overlooks entry barriers to new firms and point out that progress in eliminating racial discrimination has been slow. Discrimination based on race has persisted in the United States and other market economies decade after decade. To explain why, economists have proposed alternative models.

Statistical Discrimination

A second theory of discrimination centers on the concept of **statistical discrimination,** in which people are judged on the basis of the average characteristics of the group to which they belong, rather than on their own personal characteristics or productivity. For example, insurance rates for teenage males are higher than those for teenage females. The difference is based on factual evidence indicating that, on average, young males are more likely than young females to be in accidents. But many young men are actually less accident-prone than the average young woman, and those men are discriminated against by having to pay higher insurance rates. The uniqueness of the theory of statistical discrimination is its suggestion that discriminatory outcomes are possible even where there is no prejudice.

Labor Market Example How does statistical discrimination show itself in labor markets? Employers with job openings want to hire the most productive workers available. They have their personnel department collect information concerning each job applicant, including age, education, and prior work experience. They may supplement that information with preemployment tests, which they feel are helpful indicators of potential job performance. But collecting detailed information about job applicants is very expensive, and predicting job performance on the basis of limited data is difficult. Consequently, some employers looking for inexpensive information may consider the *average* characteristics of women and minorities in determining whom to hire. They are in fact practicing statistical discrimination when they do so. They are using gender, race, or ethnic background as a crude indicator of production-related attributes.

Example: Suppose an employer who plans to invest heavily in training a worker knows that on average women are less likely to be career-oriented than men, more likely to quit work in order to care for young children, and more likely to refuse geographic transfers. Thus, on average, the return on the employer's investment in training is likely to be less when choosing a woman than when choosing a man. All else equal, when choosing between two job applicants, one a woman and the other a man, this employer is likely to hire the man.

Note what is happening here. Average characteristics for a *group* are being applied to *individual* members of that group. The employer is falsely assuming that *each and every* woman worker has the same employment tendencies as the *average* woman. Such stereotyping means that numerous women who are career-oriented, who do not plan on quitting work in order to care for young children, and who are flexible as to geographic transfers will be discriminated against.

Profitable, Undesirable, but Not Malicious The firm that practices statistical discrimination is not being malicious in its hiring behavior (although it may be violating antidiscrimination laws). The decisions it makes will be rational and profitable because *on average* its hiring decisions are likely to be correct. Nevertheless, many people suffer because of statistical discrimination, since it blocks the economic betterment of capable people. And since it is profitable, statistical discrimination tends to persist.

Occupational Segregation: The Crowding Model

The practice of **occupational segregation**—the crowding of women, African Americans, and certain ethnic groups into less desirable, lower-paying occupations—is still apparent in the U.S. economy. Statistics indicate that women are disproportionately concentrated in a limited number of occupations such as teaching, nursing, and secretarial and clerical jobs. African Americans and Hispanics are crowded into low-paying jobs such as those of laundry workers, cleaners and household aides, hospital orderlies, agricultural workers, and other manual laborers.

Let's look at a model of occupational segregation, using women and men as an example.

The Model The character and income consequences of occupational discrimination are revealed through a labor supply and demand model. We make the following assumptions:

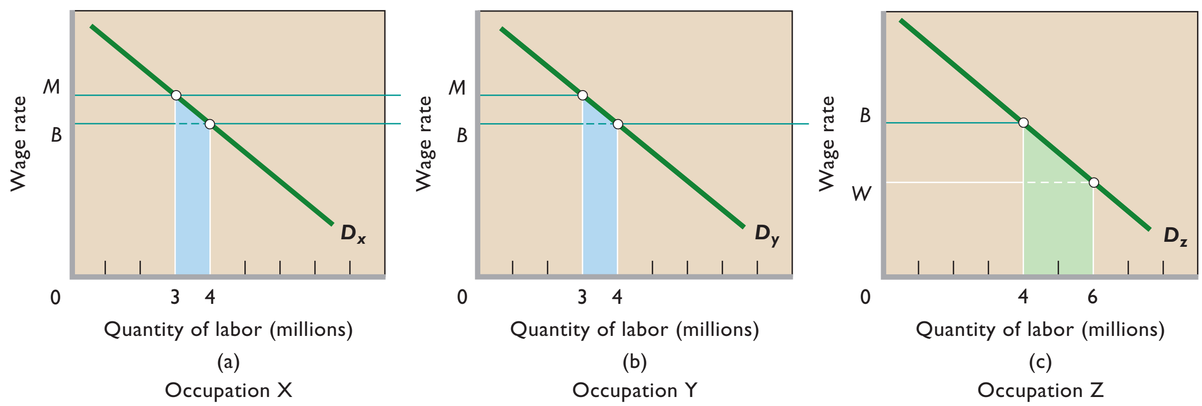
- The labor force is equally divided between men and women workers. Let's say there are 6 million male and 6 million female workers.
- The economy comprises three occupations, X, Y, and Z, with identical labor demand curves, as shown in Figure 21.7.
- Men and women have the same labor-force characteristics; each of the three occupations could be filled equally well by men or by women.

Effects of Crowding Suppose that, as a consequence of discrimination, the 6 million women are excluded from occupations X and Y and crowded into occupation Z, where they earn wage *W*. The men distribute themselves equally among occupations X and Y, meaning that 3 million male workers are in each occupation and have a common wage of *M*. (If we assume that there are no barriers to mobility between X and Y, any initially different distribution of males between X and Y would result in a wage differential between the two occupations. That would prompt labor shifts from the low- to the high-wage occupation until an equal distribution occurred.)

Because women are crowded into occupation Z, labor supply (not shown) is larger and their wage rate *W* is much lower than *M*. Because of the discrimination, this is an equilibrium situation that will persist as long as the crowding occurs. The occupational barrier means women cannot move into occupations X and Y in pursuit of a higher wage.

The result is a loss of output for society. To see why, recall again that labor demand reflects labor's marginal revenue product, which is labor's contribution to domestic output. Thus, the blue areas for occupations X and Y in Figure 21.7 show the decrease in domestic output—the market value of the marginal output—caused by subtracting 1 million women from each of these occupations. Similarly, the green area for occupation Z shows the

FIGURE 21.7 The economics of occupational segregation. (a) Because 1 million women are excluded from Occupation X, the wage rate for men in that occupation is *M* rather than *B*. (b) Because another 1 million women are excluded from Occupation Y, the wage for men in Occupation Y is also *M* rather than *B*. (c) Because the 2 million women excluded from Occupations X and Y are crowded into Occupation Z, the wage rate for women there is *W* rather *B*. The elimination of discrimination will create flows of 1 million women to Occupation X and 1 million women to Occupation Y. The wage rate will equalize at *B* in all three occupations and the nation's output will rise by the sum of the two blue areas minus the single green area.



increase in domestic output caused by moving 2 million women into occupation Z. Although society would gain the added output represented by the green area in occupation Z, it would lose the output represented by the sum of the two blue areas in occupations X and Y. That output loss exceeds the output gain, producing a net output loss for society.

Eliminating Occupational Segregation Now assume that through legislation or sweeping changes in social attitudes, discrimination disappears. Women, attracted by higher wage rates, shift from occupation Z to X and Y; 1 million women move into X and another 1 million move into Y. Now there are 4 million workers in Z, and occupational segregation is eliminated. At that point there are 4 million workers in each occupation, and wage rates in all three occupations are equal, here at *B*. That wage equality eliminates the incentive for further reallocations of labor.

The new, nondiscriminatory equilibrium clearly benefits women, who now receive higher wages; it hurts men, who now receive lower wages. But women were initially harmed and men benefited through discrimination; removing discrimination corrects that situation.

Society also gains. The elimination of occupational segregation reverses the net output loss just discussed. Adding 1 million women to each of occupations X and Y in Figure 21.7 increases domestic output by the sum of the two blue areas. The decrease in domestic output caused by losing 2 million women from occupation Z is shown by the green area. The sum of the two increases in domestic output in X and Y exceeds the decrease in domestic output in Z. With the end of the discrimination, 2 million women workers have moved from occupation Z, where their contribution to domestic output (their MRP) is low, to higher-paying occupations X and Y, where their contribution to domestic output is high. Thus society gains a more efficient allocation of resources from the removal of occupational discrimination.

Example: The easing of occupational barriers has led to a surge of women gaining advanced degrees in some high-paying professions. In recent years, for instance, the percentage of law degrees and medical degrees awarded to women has exceeded 50 percent, compared with less than 10 percent in 1970.

QUICK REVIEW 21.3

- Discrimination occurs when workers who have the same abilities, education, training, and experience as other workers receive inferior treatment with respect to hiring, occupational access, promotion, or wages.
- The taste-for-discrimination model sees discrimination as representing a preference or "taste" for which the discriminator is willing to pay.
- The theory of statistical discrimination says that employers often wrongly judge individuals on the basis of average group characteristics rather than on personal characteristics, thus harming those discriminated against.
- The crowding model of discrimination suggests that when women and minorities are systematically excluded from high-paying occupations and crowded into low-paying ones, their wages and society's domestic output are reduced.

LAST WORD

U.S. Family Wealth and Its Distribution

Between 1995 and 2010, Family Wealth Rose Rapidly for Both the Typical Family as Well as Very Well-Off Families. But Then the Collapse of the Housing Bubble and the Recession of 2007–2009 Greatly Reduced the Typical Family's Level of Wealth.

The latest findings on U.S. family wealth (= net worth = assets *minus* liabilities) were reported in 2012. The data were obtained from the Survey of Consumer Finances and *The State of Working America*. The information includes data on median family wealth, average family wealth, and the distribution of wealth in the United States. *Median family wealth* is the wealth received by the family at the midpoint of the wealth distribution; *average family wealth* is simply total wealth divided by the number of families.

As shown in Table 1, median and average family wealth, adjusted for inflation, both grew rapidly between 1995 and 2007. But then the 2007–2009 recession caused both to drop precipitously. Median wealth, in particular, fell so far that its 2010 value of $77,300 was nearly 8 percent lower than the $84,000 value observed 15 years earlier.

Most of the decline was due to the 30 percent fall in average housing prices that took place during and after the recession. Millions of people who had taken out loans to buy houses found themselves "underwater" on their home mortgages because the market values of their homes had fallen below what they owed on their loans. That meant that they had *negative* wealth in terms of their housing purchases. As a result, overall wealth levels dropped sharply.

Table 2 looks at the distribution of family wealth for various percentile groups and reveals that the distribution of wealth is highly unequal. In 2010 the wealthiest 10 percent of families owned almost 77 percent of the total wealth and the top 1 percent owned about 35 percent. By contrast, the bottom 90 percent held only about 23 percent of the total wealth. Moreover, the general trend is toward greater inequality. The lowest 90 percent

of families saw their share fall from about 32 percent of total U.S. wealth in 1995 to only about 23 percent in 2010.

Tables 1 and 2 raise many interesting questions: Will the inequality of wealth continue to grow in the future? If so, what are the implications for the future character of American society? Should government do more, or less, in the future to try to redistribute wealth? Would new government policies to redistribute wealth endanger or slow the creation of wealth and the growth of income for average Americans? As of 2013, the federal estate (inheritance) tax had a zero percent rate up to $5.25 million and then a 40 percent rate on anything above $5.25 million. There is a vigorous debate about whether this represents too much or too little in the way of estate taxation.

Sources: "Changes in U.S. Family Finances from 2004 to 2007: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, vol. 95 (February 2009); "Ponds and Streams: Wealth and Income in the U.S., 1989 to 2007," Survey of Consumer Finances working paper, January 2009, p. 35; Lawrence Mishel, Josh Bivens, Elise Gould, and Heidi Shierholz, *The State of Working America, 12th ed.*, Economic Policy Institute (2012).

TABLE 1 Median and Average Family Wealth, Survey Years 1995–2010 (in 2010 Dollars)

Year	Median	Average*
1995	$ 84,000	$307,900
1998	98,100	386,700
2001	106,100	487,000
2004	107,200	517,100
2007	126,400	584,600
2010	77,300	498,800

*The averages greatly exceed the medians because the averages are boosted by the multibillion-dollar wealth of a relatively few families.

TABLE 2 Percentage of Total Family Wealth Held by Different Groups, Survey Years 1995–2010

	Percentage of Total Wealth by Group		
Year	Bottom 90%	Top 10%	Top 1%
1995	32.2%	67.8%	34.6%
1998	31.4	68.6	33.9
2001	30.2	69.8	32.7
2004	30.4	69.5	33.4
2007	28.5	71.5	33.8
2010	23.3	76.7	35.4

Cost to Society as Well as to Individuals

It is obvious from all three models of discrimination that discrimination by characteristics such as race, ethnicity, gender, or age imposes costs on those who are discriminated against. They have lower wages, less access to jobs, or both. Preferred workers in turn benefit from discrimination through less job competition, greater job access, and higher wages. But discrimination does more than simply transfer earnings from some people to others, thus contributing to income inequality and increasing poverty. Where it exists, discrimination also diminishes the economy's total output and income. In that regard, discrimination acts much like any other artificial barrier to free competition. By arbitrarily blocking qualified individuals from high-productivity (and thus high-wage) jobs, discrimination keeps those discriminated against from providing their maximum contribution to society's total output and total income. In terms of production possibilities analysis, discrimination locates society inside the production possibilities curve that would be available to it if there were no discrimination.

Discrimination redistributes a diminished amount of total income.

SUMMARY

LO21.1 Explain how income inequality in the United States is measured and described.

The distribution of income in the United States reflects considerable inequality. The richest 20 percent of households receive 51.1 percent of total income, while the poorest 20 percent receive 3.2 percent.

The Lorenz curve shows the percentage of total income received by each percentage of households. The extent of the gap between the Lorenz curve and a line of total equality illustrates the degree of income inequality.

The Gini ratio measures the overall dispersion of the income distribution and is found by dividing the area between the diagonal and the Lorenz curve by the entire area below the diagonal. The Gini ratio ranges from zero to 1, with higher ratios signifying greater degrees of income inequality.

LO21.2 Discuss the extent and sources of income inequality.

Recognizing that the positions of individual families in the distribution of income change over time and incorporating the effects of noncash transfers and taxes would reveal less income inequality than do standard census data. Government transfers (cash and noncash) greatly lessen the degree of income inequality; taxes also reduce inequality, but not nearly as much as transfers.

Causes of income inequality include differences in abilities, in education and training, and in job tastes, along with discrimination, inequality in the distribution of wealth, and an unequal distribution of market power.

LO21.3 Demonstrate how income inequality has changed since 1975.

Census data show that income inequality has increased since 1975. The major cause of the recent increases in income inequality is a rising demand for highly skilled workers, which has boosted their earnings significantly.

LO21.4 Debate the economic arguments for and against income inequality.

The basic argument for income equality is that it maximizes consumer satisfaction (total utility) from a particular level of total income. The main argument for income inequality is that it provides the incentives to work, invest, and assume risk and is necessary for the production of output, which, in turn, creates income that is then available for distribution.

LO21.5 Relate how poverty is measured and its incidence by age, gender, ethnicity, and other characteristics.

Current statistics reveal that 15 percent of the U.S. population lives in poverty. Poverty rates are particularly high for female-headed families, young children, African Americans, and Hispanics.

LO21.6 Identify the major components of the income-maintenance program in the United States.

The present income-maintenance program in the United States consists of social insurance programs (Social Security, Medicare, and unemployment compensation) and public assistance programs (SSI, TANF, food stamps, Medicaid, and the earned-income tax credit).

LO21.7 Discuss labor market discrimination and how it might affect hiring decisions and wages.

Discrimination relating to the labor market occurs when women or minorities having the same abilities, education, training, and experience as men or white workers are given inferior treatment with respect to hiring, occupational choice, education and training, promotion, and wage rates. Discrimination redistributes national income and, by creating inefficiencies, diminishes its size.

In the taste-for-discrimination model, some white employers have a preference for discrimination, measured by a discrimination coefficient *d*. Prejudiced white employers will hire African-American workers only if their wages are at least *d* dollars below those of whites. The model indicates that declines in the discrimination coefficients of white employers will increase the demand for African-American workers, raising the African-American wage rate and the ratio of African-American wages to white wages. It also suggests that competition may eliminate discrimination in the long run.

Statistical discrimination occurs when employers base employment decisions about *individuals* on the average characteristics of *groups* of workers. That can lead to discrimination against individuals even in the absence of prejudice.

The crowding model of occupational segregation indicates how white males gain higher earnings at the expense of women and certain minorities who are confined to a limited number of occupations. The model shows that discrimination also causes a net loss of domestic output.

TERMS AND CONCEPTS

income inequality
Lorenz curve
Gini ratio
income mobility
noncash transfers
equality-efficiency trade-off
poverty rate
entitlement programs
social insurance programs
Social Security
Medicare
unemployment compensation
public assistance programs
Supplemental Security Income (SSI)
Temporary Assistance for Needy Families (TANF)
Supplemental Nutrition Assistance Program (SNAP)
Medicaid
earned-income tax credit (EITC)
discrimination
taste-for-discrimination model
discrimination coefficient
statistical discrimination
occupational segregation

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Use quintiles to briefly summarize the degree of income inequality in the United States. How and to what extent does government reduce income inequality? **LO21.1**
2. Assume that Al, Beth, Carol, David, and Ed receive incomes of \$500, \$250, \$125, \$75, and \$50, respectively. Construct and interpret a Lorenz curve for this five-person economy. What percentage of total income is received by the richest quintile and by the poorest quintile? **LO21.1**
3. How does the Gini ratio relate to the Lorenz curve? Why can't the Gini ratio exceed 1? What is implied about the direction of income inequality if the Gini ratio declines from 0.42 to 0.35? How would one show that change of inequality in the Lorenz diagram? **LO21.1**
4. Why is the lifetime distribution of income more equal than the distribution in any specific year? **LO21.1**
5. Briefly discuss the major causes of income inequality. With respect to income inequality, is there any difference between inheriting property and inheriting a high IQ? Explain. **LO21.2**
6. What factors have contributed to increased income inequality since 1975? **LO21.3**
7. Should a nation's income be distributed to its members according to their contributions to the production of that total income or according to the members' needs? Should society attempt to equalize income or economic opportunities? Are the issues of equity and equality in the distribution of income synonymous? To what degree, if any, is income inequality equitable? **LO21.4**
8. Do you agree or disagree? Explain your reasoning. "There need be no trade-off between equality and efficiency. An 'efficient' economy that yields an income distribution that many regard as unfair may cause those with meager incomes to become discouraged and stop trying. So efficiency may be undermined. A fairer distribution of rewards may generate a higher average productive effort on the part of the population, thereby enhancing efficiency. If people think they are playing a fair economic game and this belief causes them to try harder, an economy with an equitable income distribution may be efficient as well."[6] **LO21.4**
9. Comment on or explain: **LO21.4**
   a. Endowing everyone with equal income will make for very unequal enjoyment and satisfaction.

[6]Paraphrased from Andrew Schotter, *Free Market Economics* (New York: St. Martin's Press, 1985), pp. 30–31.

b. Equality is a "superior good"; the richer we become, the more of it we can afford.
c. The mob goes in search of bread, and the means it employs is generally to wreck the bakeries.
d. Some freedoms may be more important in the long run than freedom from want on the part of every individual.
e. Capitalism and democracy are really a most improbable mixture. Maybe that is why they need each other—to put some rationality into equality and some humanity into efficiency.
f. The incentives created by the attempt to bring about a more equal distribution of income are in conflict with the incentives needed to generate increased income.

10. How do government statisticians determine the poverty rate? How could the poverty rate fall while the number of people in poverty rises? Which group in each of the following pairs has the higher poverty rate: (*a*) children or people age 65 or over? (*b*) African Americans or foreign-born noncitizens? (*c*) Asians or Hispanics? **LO21.5**

11. What are the essential differences between social insurance and public assistance programs? Why is Medicare a social insurance program, whereas Medicaid is a public assistance program? Why is the earned-income tax credit considered to be a public assistance program? **LO21.6**

12. The labor demand and supply data in the following table relate to a single occupation. Use them to answer the questions that follow. Base your answers on the taste-for-discrimination model. **LO21.7**

Quantity of Hispanic Labor Demanded, Thousands	Hispanic Wage Rate	Quantity of Hispanic Labor Supplied, Thousands
24	$16	52
30	14	44
35	12	35
42	10	28
48	8	20

a. Plot the labor demand and supply curves for Hispanic workers in this occupation.
b. What are the equilibrium Hispanic wage rate and quantity of Hispanic employment?
c. Suppose the white wage rate in this occupation is $16. What is the Hispanic-to-white wage ratio?
d. Suppose a particular employer has a discrimination coefficient *d* of $5 per hour. Will that employer hire Hispanic or white workers at the Hispanic-white wage ratio indicated in part *c*? Explain.
e. Suppose employers as a group become less prejudiced against Hispanics and demand 14 more units of Hispanic labor at each Hispanic wage rate in the table. What are the new equilibrium Hispanic wage rate and level of Hispanic employment? Does the Hispanic-white wage ratio rise or fall? Explain.
f. Suppose Hispanics as a group increase their labor services in that occupation, collectively offering 14 more units of labor at each Hispanic wage rate. Disregarding the changes indicated in part *e*, what are the new equilibrium Hispanic wage rate and level of Hispanic employment? Does the Hispanic-white wage ratio rise, or does it fall?

13. Males under the age of 25 must pay far higher auto insurance premiums than females in this age group. How does this fact relate to statistical discrimination? Statistical discrimination implies that discrimination can persist indefinitely, while the taste-for-discrimination model suggests that competition might reduce discrimination in the long run. Explain the difference. **LO21.7**

14. Use a demand-and-supply model to explain the impact of occupational segregation or "crowding" on the relative wage rates and earnings of men and women. Who gains and who loses from the elimination of occupational segregation? Is there a net gain or a net loss to society? Explain. **LO21.7**

15. **LAST WORD** Go to Table 1 in the Last Word and compute the ratio of average wealth to median wealth for each of the 6 years. What trend do you find? What is your explanation for the trend? The federal estate tax redistributes wealth in two ways: by encouraging charitable giving, which reduces the taxable estate, and by heavily taxing extraordinarily large estates and using the proceeds to fund government programs. Do you favor repealing the estate tax? Explain.

REVIEW QUESTIONS

1. Suppose that the United States has a Gini ratio of 0.41 while Sweden has a Gini ratio of 0.31. Which country has a more equal distribution of income? **LO21.1**
a. The United States.
b. Sweden.
c. They are actually equal.

2. Some part of income inequality is likely to be the result of discrimination. But other factors responsible for inequality include (select as many as apply): **LO21.2**
a. Differences in abilities and talents.
b. Differences in education and training.
c. Different preferences for work versus leisure.
d. Different preferences for low-paying but safe jobs relative to high-paying but dangerous jobs.

3. Suppose that a society contains only two members, a lawyer named Monique and a handyman named James. Five years ago, Monique made $100,000 while James made $50,000. This year, Monique will make $300,000 while James will make $100,000. Which of the following statements about this society's income distribution are true? **LO21.2**

*Select **one or more** answers from the choices shown.*

a. In absolute dollar amounts, the entire distribution of income has been moving upward.
b. In absolute dollar amounts, the entire distribution of income has been stagnant.
c. The relative distribution of income has become more equal.
d. The relative distribution of income has become less equal.
e. The relative distribution of income has remained constant.
f. The rich are getting richer while the poor are getting poorer.
g. The rich are getting richer faster than the poor are getting richer.

4. Suppose that the last dollar that Victoria receives as income brings her a marginal utility of 10 utils while the last dollar that Fredrick receives as income brings him a marginal utility of 15 utils. If our goal is to maximize the combined total utility of Victoria and Fredrick, we should: **LO21.4**

a. Redistribute income from Victoria to Fredrick.
b. Redistribute income from Fredrick to Victoria.
c. Not engage in any redistribution because the current situation already maximizes total utility.
d. None of the above.

5. If women are crowded into elementary education and away from fire fighting, wages in fire fighting will tend to be ______________ than if women weren't crowded into elementary education. **LO21.7**

a. Higher.
b. Lower.

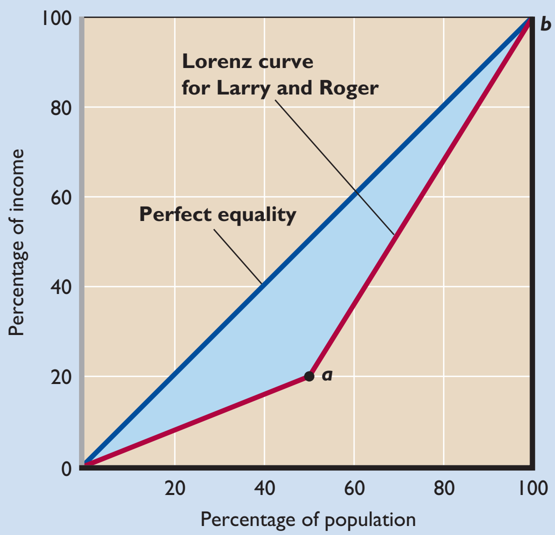
6. In the taste-for-discrimination model, an increase in employer prejudice against African-American workers would cause the discrimination coefficient to ______________ and the demand curve for African-American labor to shift ______________. **LO21.7**

a. Decrease; right.
b. Decrease; left.
c. Increase; right.
d. Increase; left.

PROBLEMS

1. In 2012 *Forbes* magazine listed Bill Gates, the founder of Microsoft, as the richest person in the United States. His personal wealth was estimated to be $66 billion. Given that there were about 309 million people living in the United States that year, how much could each person have received if Gates' wealth had been divided equally among the population of the United States? (Hint: A billion is a 1 followed by 9 zeros while a million is a 1 followed by six zeros.) **LO21.1**

2. Imagine an economy with only two people. Larry earns $20,000 per year, while Roger earns $80,000 per year. As shown in the following figure, the Lorenz curve for this two-person economy consists of two line segments. The first runs from the origin to point *a*, while the second runs from point *a* to point *b*. **LO21.1**



a. Calculate the Gini ratio for this two-person economy using the geometric formulas for the area of a triangle (= ½ × base × height) and the area of a rectangle (= base × height). (Hint: The area under the line segment from point *a* to point *b* can be thought of as the sum of the area of a particular triangle and the area of a particular rectangle.)
b. What would the Gini ratio be if the government taxed $20,000 away from Roger and gave it to Larry? (Hint: The figure will change.)
c. Start again with Larry earning $20,000 per year and Roger earning $80,000 per year. What would the Gini ratio be if both their incomes doubled? How much has the Gini ratio changed from before the doubling in incomes to after the doubling in incomes?

3. In 2012, many unskilled workers in the United States earned the federal minimum wage of $7.25 per hour. By contrast, average earnings in 2012 were about $22 per hour, and certain highly skilled professionals, such as doctors and lawyers, earned $100 or more per hour. **LO21.6**

a. If we assume that wage differences are caused solely by differences in productivity, how many times more productive was the average worker than a worker being paid the federal minimum wage? How many times more productive was a $100-per-hour lawyer compared to a worker earning minimum wage?
b. Assume that there are 20 minimum-wage workers in the economy for each $100-per-hour lawyer. Also assume that both lawyers and minimum-wage workers work the same number of hours per week. If everyone works 40 hours per week, how much does a $100-per-hour lawyer earn a week? How much does a minimum-wage worker earn a week?

   c. Suppose that the government pairs each $100-per-hour lawyer with 20 nearby minimum-wage workers. If the government taxes 25 percent of each lawyer's income each week and distributes it equally among the 20 minimum-wage workers with whom each lawyer is paired, how much will each of those minimum-wage workers receive each week? If we divide by the number of hours worked each week, how much does each minimum-wage worker's weekly transfer amount to on an hourly basis?
   d. What if instead the government taxed each lawyer 100 percent before dividing the money equally among the 20 minimum-wage workers with whom each lawyer is paired—how much per week will each minimum-wage worker receive? And how much is that on an hourly basis?

4. The desire to maximize profits can work against racial and other types of discrimination. To see this, consider two equally productive accountants named Ted and Jared. Ted is black, and Jared is white. Both can complete 10 audits per month. **LO21.7**
   a. Suppose that for any accounting firm that hires either Ted or Jared, all the other costs of performing an audit (besides paying either Ted or Jared) come to $1,000 per audit. If the going rate that must be paid to hire an accountant is $7,000 per month, how much will it cost an accounting firm to produce one audit if it hires either Ted or Jared to do the work?
   b. If the market price that accounting firms charge their clients for an audit is $1,800, what would the accounting profit per audit be for a firm that hired either Ted or Jared? What is the profit rate as a percentage?
   c. Suppose that firm *A* dislikes hiring black accountants, while firm *B* is happy to hire them. So Ted ends up working at firm *B* rather than firm *A*. If Ted works 11 months per year, how many audits will he complete for firm *B* each year? How much in accounting profits will firm *B* earn each year from those audits?
   d. Because firm *A* passed on hiring Ted because he was black, firm *A* is forgoing the profits it could have earned if it had hired Ted. If the firm is willing to forgo up to $5,000 per year in profit to avoid hiring blacks, by how many dollars will firm *A* regret its decision not to hire Ted?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **22**

Health Care

Learning Objectives

LO22.1 **Convey important facts about rising health care costs in the United States.**

LO22.2 **Relate the economic implications of rising health care costs.**

LO22.3 **Discuss the problem of limited access to health care for those without insurance.**

LO22.4 **List the demand and supply factors explaining rising health care costs.**

LO22.5 **Describe the cost-containment strategies that rely on altering the financial incentives facing either patients or health service providers.**

LO22.6 **Summarize the goals of the Patient Protection and Affordable Care Act and the major changes that it institutes.**

On March 23, 2010, President Barack Obama signed into law the **Patient Protection and Affordable Care Act (PPACA),** a wide-ranging law that proponents claimed would lower health care costs while increasing access to quality health care for millions of poorer Americans.

At over 2,400 pages, the legislation was designed to address a wide set of concerns relating to the provision, delivery, and cost of health care. These included the high and rapidly rising cost of health insurance for those who did have health insurance, the fact that tens of millions of Americans at any given moment were without health insurance, and the inability of many people with preexisting conditions to obtain health insurance.

The controversial law gave the federal government sweeping new powers to promote universal insurance coverage and to regulate the details of insurance policies. Because health care spending was 17.9 percent of GDP in 2010, the law effectively put the federal government in control of nearly one-fifth of the U.S. economy. This chapter applies microeconomic analysis to help explain the origin of the problems that the law was designed to address as well as the heated debate over whether the policies prescribed by the law are likely to achieve their goals.

The Health Care Industry

LO22.1 Convey important facts about rising health care costs in the United States.

Because the boundaries of the health care industry are not precise, defining the industry is difficult. In general, it includes services provided in hospitals, nursing homes, laboratories, and physicians' and dentists' offices. It also includes prescription and nonprescription drugs, artificial limbs, and eyeglasses. Note, however, that many goods and services that may affect health are not included, for example, low-fat foods, vitamins, and health club services.

Health care is one of the largest U.S. industries, employing about 17 million people, including about 850,000 practicing physicians, or 275 doctors per 100,000 of population. There are about 5,800 hospitals containing 925,000 beds. Americans make more than 1 billion visits to office-based physicians each year.

The U.S. Emphasis on Private Health Insurance

Many of the provisions of the Patient Protection and Affordable Care Act are focused on health insurance. This is because a high proportion of health care spending in the United States is provided through private health insurance policies paid for by employers. By contrast, many countries such as Canada have systems of **national health insurance** in which the government uses tax revenues to provide a basic package of health care to every resident at either no charge or at low cost-sharing levels. In such countries only a relatively few people bother to buy private health insurance—and then only to cover services that are not paid for by the national health insurance system.

The uniquely American emphasis on private health insurance paid for by employers is a relatively recent phenomenon. It began during the Second World War in response to price and wage controls that the federal government imposed to prevent inflation. The wage controls were problematic for the private companies charged with building the tanks, planes, and boats needed to win the war. These firms needed to expand output rapidly and knew that doing so would be possible only if they could attract workers away from other industries. Several manufacturers stumbled on the strategy of offering free health insurance as a way of attracting workers. Unable to raise wages, the companies recruited the workers they needed by offering health insurance as a fringe benefit paid for by the employer.

After the war, price and wage controls were lifted. Nevertheless, more and more companies began to offer "free" health insurance to their employees. They did so because of a provision in the federal tax law that makes it cheaper for companies to purchase insurance for their employees than it would be for employees to purchase insurance on their own behalf. By 2007, this incentive structure led to a situation in which nearly 88 percent of people with private health insurance received it as a benefit provided by their employer rather than by purchasing it themselves directly from an insurance company.

The prominence of employer-provided health insurance in the United States has had several important consequences. Perhaps the most important is that health care paid for via insurance can create perverse incentives for overuse that, in turn, lead to higher prices.

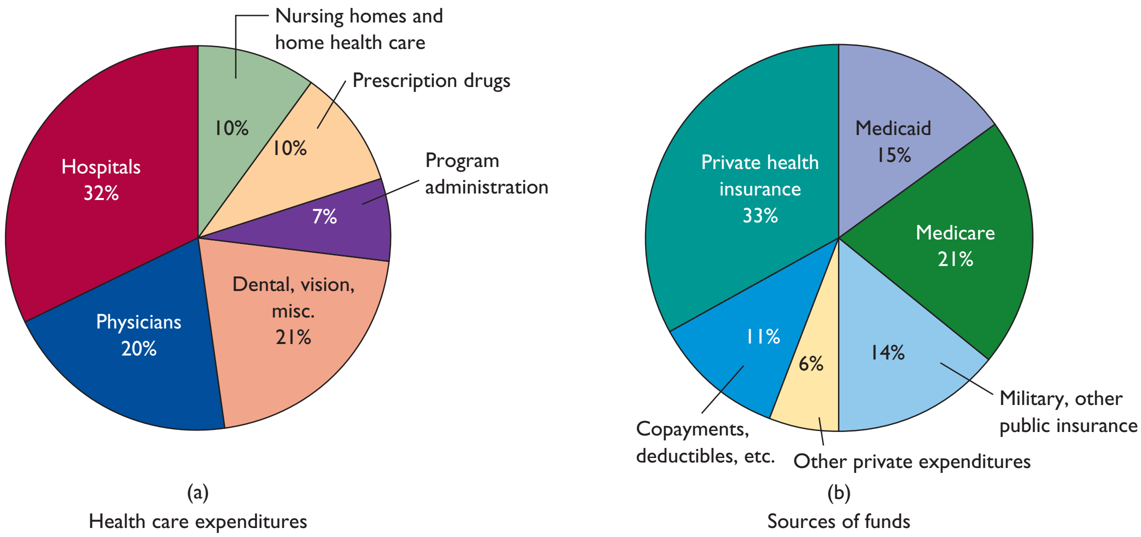
Another consequence of employer-provided health care is that health care reform efforts have focused on regulating the health insurance system with which most people are familiar rather than attempting alternatives that most people have never experienced. Later, we will explore how this tendency to regulate—rather than replace—the current insurance system has affected recent reform efforts including the Patient Protection and Affordable Care Act.

Twin Problems: Costs and Access

In recent decades, the U.S. health care system has suffered from two highly publicized problems:

- The cost of health care has risen rapidly in response to higher prices and an increase in the quantity of services provided. (Spending on health care involves both "prices" and "quantities" and is often loosely referred to as "health care costs.") The price of medical care has traditionally increased faster than the overall price level in nearly all periods. A pleasant exception happened in the late 2000s. The December-to-December index of medical care prices rose by 2.6 percent in 2008, 3.4 percent in 2009, 3.4 percent in 2010, and 3.0 percent in 2011,

FIGURE 22.1 Health care expenditures and finance. Total U.S. health care expenditures are extremely large ($2.7 trillion in 2011). (a) Most health care expenditures are for hospitals and the services of physicians and other skilled professionals. (b) Public and private insurance pay for four-fifths of health care.



Source: Centers for Medicare and Medicaid Services, **cms.hhs.gov**. Data are for 2011 and are compiled by the authors.

for a four-year average of 3.1 percent. At the same time, the overall price index increased by an annual average of 3.1 percent for those four years. However, health care spending (price × quantity) grew by 6.2 percent in 2007, 4.7 percent in 2008, 3.8 percent in 2009, and 3.9 percent in 2010. It is projected to grow at an annual rate of 5.8 percent over the next 10 years, far faster than inflation.

- Some 49 million Americans in 2011 did not have health insurance coverage and, as a result, had significantly reduced access to quality health care.

Efforts to reform health care have focused on controlling costs and making it accessible to everyone. Those two goals are related, since high and rising prices make health care services unaffordable to a significant portion of the U.S. population. In fact, a dual system of health care may be evolving in the United States. Those with insurance or other financial means receive world-class medical treatment, but many people, because of their inability to pay, often fail to seek out even the most basic treatment. When they do seek treatment, they may receive poorer care than those who have insurance. Free county hospitals and private charity hospitals do provide services to those without insurance, but the quality of care can be considerably lower than that available to people who have insurance.

High and Rising Health Care Costs

We need to examine several aspects of health care costs, or, alternatively, health care spending.

Health Care Spending Health care spending in the United States is high and rising in both absolute terms and as a percentage of domestic output.

Total Spending on Health Care Figure 22.1a gives an overview of the major types of U.S. health care spending ($2.7 trillion in 2011). It shows that 32 cents of each health care dollar goes to hospitals, while 20 cents goes to physicians, and 21 cents is spent on dental, vision, and other miscellaneous health care services.

Figure 22.1b shows the sources of funds for health care spending. It reveals that four-fifths of health care spending is financed by insurance. Public insurance (Medicaid, Medicare, and insurance for veterans, current military personnel, and government employees) is the source of 50 cents of each dollar spent. Private insurance accounts for 33 cents. So public and private insurance combined provide 83 cents of each dollar spent. The remaining 17 cents comes directly out of the health care consumer's pocket. It is paid mainly as insurance **deductibles** (that is, the insured pays the first $250 or $500 of each year's health care costs before the insurer begins paying) or **copayments** (that is, the insured pays, say, 20 percent of all health care costs and the insurance company pays 80 percent).

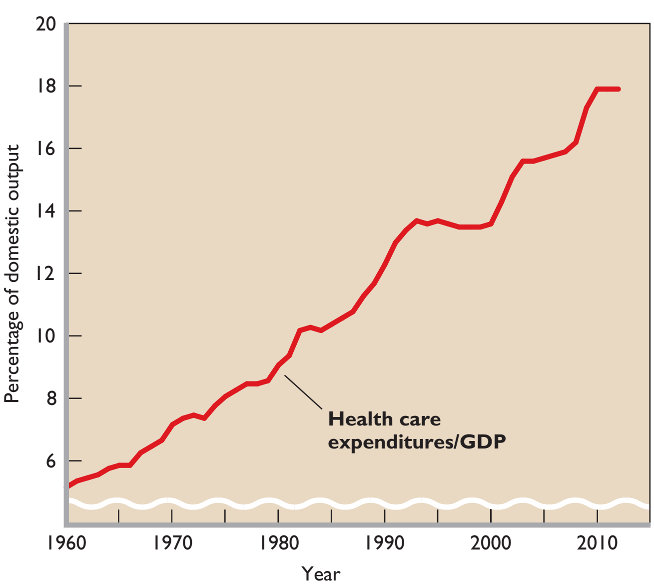
As we discuss in Chapter 21, Medicare is a nationwide federal health care program available to Social Security beneficiaries and persons with disabilities. One part of Medicare is a hospital insurance program that, after a deductible of $1,100 in 2010, covers all reasonable costs for the first 60 days of inpatient care per "benefit period" and lesser amounts (on a cost-sharing basis) for additional days. Coverage is also provided for posthospital nursing services, home health care, and hospice care for the terminally ill. Other parts of Medicare (including insurance programs for physicians' services, laboratory and other diagnostic tests, outpatient hospital services, and prescription drugs) are voluntary but heavily subsidized by government. The monthly premiums that most participants pay cover about one-fourth of the cost of the benefits provided.

Medicaid provides payment for medical benefits to certain low-income people, including the elderly, the blind, persons with disabilities, children, and adults with dependent children. Those who qualify for Temporary Aid for Needy Families (TANF) and the Supplemental Security Income (SSI) program are automatically eligible for Medicaid. Nevertheless, Medicaid covers less than half of those living in poverty. The federal government and the states share the cost of Medicaid. On average, the states fund 42 percent and the federal government 58 percent of each Medicaid dollar spent.

Overall, about 17 percent of each dollar spent on health care is financed by direct out-of-pocket payments by individuals. The fact that most U.S. health care is paid for by private insurance companies or the government is an important contributor to rising health care costs.

Percentage of GDP Figure 22.2 shows how U.S. health care spending has been increasing as a percentage of GDP.

FIGURE 22.2 U.S. health care expenditures as a percentage of GDP. U.S. health care spending as a percentage of GDP has greatly increased since 1960.

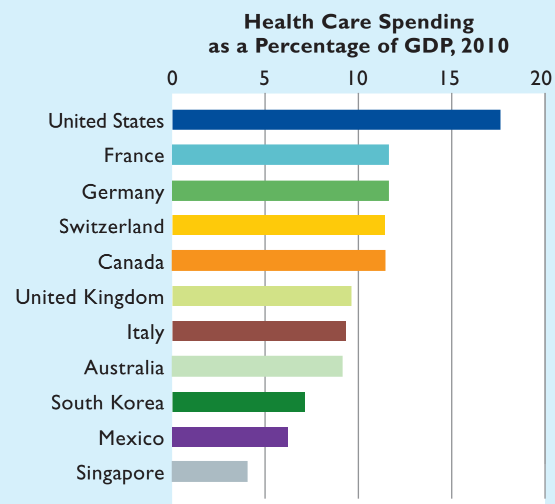


Source: Centers for Medicare and Medicaid Services, **cms.hhs.gov**.

GLOBAL PERSPECTIVE 22.1

Health Care Spending as a Percentage of GDP, Selected Nations

The United States tops the chart when it comes to health care expenditures as a percentage of GDP.



Source: Organization for Economic Cooperation and Development, **www.oecd.org**.

Health care spending absorbed 5.2 percent of GDP in 1960 but rose to 17.9 percent in 2011.

International Comparisons Global Perspective 22.1 reveals that among the industrialized nations, health care spending as a percentage of GDP is highest in the United States. It is reasonable to assume that health care spending varies positively with output and incomes, but that doesn't account for the higher U.S. health expenditures as a percentage of GDP. Later in the chapter we discuss various explanations for why the United States is "in a league of its own" as to its proportion of output devoted to health care.

Quality of Care: Are We Healthier?

To compare the quality of health care from country to country is difficult. Yet there is general agreement that medical care (although not health and not "preventive treatment") in the United States is among the best in the world. Average life expectancy in the United States has increased by about 7 years since 1970, and U.S. physicians and hospitals employ the most advanced medical equipment and technologies. Also, more than half the world's medical research funding is done in the United States. As a result, the incidence of disease has been declining and the quality of treatment has been improving. Polio has been virtually eliminated, ulcers are successfully treated without surgery, angioplasty and coronary bypass surgery greatly benefit those with heart disease, sophisticated body scanners are increasingly available diagnostic tools, and organ transplants and prosthetic joint replacements are more and more common.

That is the good news. But there is other news as well. Despite new screening and treatment technologies, the breast cancer mortality rate has shown only modest improvement. Tuberculosis, a virtually forgotten disease, has reappeared. The AIDS epidemic has claimed more than 619,400 American lives. More generally, some experts say that high levels of health care spending have not produced significantly better health and well-being. U.S. health care expenditures are the highest in the world absolutely, as a proportion of GDP, and on a per capita basis. Yet many nations have lower rates of maternal and infant mortality and longer life expectancies.

Economic Implications of Rising Costs

LO22.2 Relate the economic implications of rising health care costs.

The most visible economic effects of rising health care costs are higher health insurance premiums to employers and higher out-of-pocket costs to workers. But rising health care costs have other economic effects as well.

Reduced Access to Care

Higher health care costs and insurance premiums reduce access to health care. Some employers reduce or eliminate health insurance as part of their pay packages and some uninsured workers go without private health insurance. Consequently, the number of uninsured grows. We will consider this issue in detail momentarily.

Labor Market Effects

Surging health care costs have three main effects on labor markets:

- ***Slower wage growth*** First, gains in workers' total compensation (wages plus fringe benefits, including health insurance paid for by employers) generally match gains in productivity. When health care costs (and thus insurance prices) rise more rapidly than productivity, firms wanting to maintain the existing level of health care benefits for their workers must reduce the growth of the wage portion of the total compensation package. Thus, in the long run, workers bear the burden of rising health care costs in the form of slower-growing wages.
- ***Use of part-time and temporary workers*** The high cost of employer-provided health insurance has led some employers to restructure their workforces: Full-time workers with health insurance benefits are employed in smaller numbers, and uninsured part-time or temporary workers are employed in greater numbers. Similarly, an employer with a generous but expensive health care plan might reduce its health insurance expense by discharging its insured lower-wage workers—janitors, gardeners, and cafeteria staff—and replacing them with workers employed by outside independent contractors that provide little or no health insurance for their employees.
- ***Outsourcing (and Offshoring)*** Burdened by rising insurance costs, some firms may find it profitable to shift part of their production to outside suppliers that may be either domestic or international. This outsourcing may lower labor costs in situations where the outside suppliers provide fewer medical benefits to their workers. Offshoring (international outsourcing) has shifted jobs to developing economies such as Mexico, India, and China. Although labor productivity in these countries is considerably lower than that in the United States, lower wages and employer-provided medical benefits may be sufficient to make offshoring profitable. Rising domestic medical expenses therefore may join a host of other factors, including shifts in comparative advantage, in encouraging this practice.

Personal Bankruptcies

Large, uninsured medical bills are one of the major causes of personal bankruptcy. Health care experts point out that medical bills are often the last to be paid because unlike other bills there is nothing to repossess, shut off, or foreclose. So medical bills sometimes build up beyond the point of a realistic means for full repayment. Even individuals who pay all bills in a timely fashion can find themselves in tremendous financial difficulty when they face large, uninsured medical bills for major operations (such as open-heart surgery) and expensive medical procedures (such as cancer treatment).

Impact on Government Budgets

The budgets of federal, state, and local governments are negatively affected by spiraling and sometimes unpredictable health care expenditures. In the past two decades, spending for health care through Medicare and Medicaid has been by far the fastest-growing segment of the federal budget. To cover those rising expenditures, the government must either raise taxes or reduce the portion of the budget used for national defense, education, environmental programs, scientific research, and other spending categories.

The states are also finding it difficult to cover their share of the Medicaid bill. Most of them have been forced to raise their tax rates and search for new sources of revenue, and many of them have had to reduce spending on nonhealth programs such as infrastructure maintenance, welfare, and education. Local governments face similar budget strains in trying to finance public health services, hospitals, and clinics.

Too Much Spending?

Increased spending on computers or houses would be a sign of prosperity, not a cause for alarm, because society is obtaining more of each. What is different about increased spending on health care? Maybe nothing, say some economists. William Nordhaus of Yale, for example, estimated that the economic value of increases in longevity over the last 100 years nearly equals the total value of the additional GDP produced during that period. According to Kevin Murphy and Robert Topel, economists at the University of Chicago, reduced mortality from heart disease alone contributes $1.5 trillion of benefits a year in the United States. That nearly equals the entire annual GDP of Canada.

While all economists agree that improved health care has greatly contributed to society's GDP and well-being, many economists think that health care expenditures in the United States are inefficiently large. The production of health care requires scarce resources such as capital in the form of hospitals and diagnostic equipment and the highly skilled labor of physicians, technicians, and nurses. The total output of health care in the United States may be so large that health care, at the margin, is worth less than the alternative goods and services these resources could otherwise have produced. The United States therefore may be consuming health care beyond the MB = MC point that defines efficiency.

If resources are overallocated to health care, society incurs an efficiency loss. Resources used excessively in health care could be used more productively to build new factories, support research and development, construct new bridges and roads, support education, improve the environment, or provide more consumer goods.

The suggested "too much of a good thing" results from peculiarities in the market for health care. We will see that the possibility of overspending arises from the way health care is financed, the asymmetry of information between consumers and providers, and the interaction of health insurance with technological progress in the industry.

Limited Access

LO22.3 Discuss the problem of limited access to health care for those without insurance.

The other health care problem is limited access. Even though there may be an overallocation of resources to health care, not all Americans can obtain the health care they need. Extrapolations from government surveys indicate that in 2011 about 49 million Americans, or roughly 16 percent of the population, had no health insurance for the entire year. As health care costs (and therefore health care insurance premiums) continue to rise, the number of uninsured could grow.

Who are the medically uninsured? As incomes rise, so does the probability of being insured. So it is no surprise that the uninsured are concentrated among the poor. Medicaid is designed to provide health care for the poor who are on welfare. But many poor people work at low or minimum-wage jobs without health care benefits, earning "too much" to qualify for Medicaid yet not enough to afford private health insurance. About half of the uninsured have a family head who works full time. Many single-parent families, African Americans, and Hispanics are uninsured simply because they are more likely to be poor.

Curiously, those with excellent health and those with the poorest health also tend to be uninsured. Many young people with excellent health simply choose not to buy health insurance. The chronically ill find it very difficult and too costly to obtain insurance because of the likelihood that they will incur substantial health care costs in the future. Because private health insurance is most frequently obtained through an employer, the unemployed are also likely to lack insurance.

Workers for smaller firms are also less likely to have health insurance. The main reason is that administrative expenses for a small firm may be 30 to 40 percent of insurance premiums, as opposed to only 10 percent for a large firm. Also, corporations can deduct health insurance premiums from income to obtain substantial tax savings. Small unincorporated businesses can deduct only part of their health insurance expenses.

Low-wage workers are also less likely to be insured. Earlier we noted that in the long run employers pass on the increasing expense of health care insurance to workers as lower wages. This option is not available to employers who are paying the minimum wage. Thus as health care insurance premiums rise, employers cut or eliminate this benefit from the compensation package for their minimum- and low-wage workers. As a result, these workers are typically uninsured.

Although many of the uninsured forgo health care, some do not. A few are able to pay for it out of pocket. Others may wait until their illness reaches a critical stage and then go to a hospital for admittance or to be treated in the emergency room. This form of treatment is more costly than if the patient had insurance and therefore had been treated earlier by a physician. It is estimated that hospitals provide about $36 billion of uncompensated ("free") health care per year. The hospitals then try to shift these costs to those who have insurance or who can pay out of pocket.

QUICK REVIEW 22.1

- Private, employer-funded health insurance plays a much larger role in the delivery of health care in the United States than it does in other countries.
- Health care spending in the United States has been increasing absolutely and as a percentage of gross domestic output.
- Rising health care costs have caused (a) more people to find health insurance unaffordable; (b) adverse labor market effects, including slower real-wage growth and increased use of part-time and temporary workers; and (c) restriction of nonhealth spending by governments.
- Rising health care spending may reflect an overallocation of resources to the health care industry.
- Approximately 16 percent of all Americans have no health insurance and, hence, inferior access to quality health care.

Why the Rapid Rise in Costs?

LO22.4 List the demand and supply factors explaining rising health care costs.

The rising prices, quantities, and costs of health care services are the result of the demand for health care increasing much more rapidly than supply. We will examine the reasons for this in some detail. But first it will be helpful to understand certain characteristics of the health care market.

Peculiarities of the Health Care Market

We know that purely competitive markets achieve both allocative and productive efficiency: The most desired products are produced in the least costly way. We also have found that many imperfectly competitive markets, perhaps aided by regulation or the threat of antitrust action, provide outcomes generally accepted as efficient. What, then, are the special features of the health care market that have contributed to rising prices and thus escalating costs to buyers?

- ***Ethical and equity considerations*** Ethical questions inevitably intervene in markets when decisions involve the quality of life, or literally life or death. Although we might not consider it immoral or unfair if a person cannot buy a Mercedes or a personal computer, society regards it as unjust for people to be denied access to basic health care or even to the best available health care. In general, society regards health care as an "entitlement" or a "right" and is reluctant to ration it solely by price and income.
- ***Asymmetric information*** Health care buyers typically have little or no understanding of complex diagnostic and treatment procedures, while the physicians, who are the health care sellers of those procedures, possess detailed information. This creates the unusual situation in which the doctor (supplier) as the agent of the patient (consumer) tells the patient what health care services he or she should consume. We will say more about this shortly.
- ***Positive externalities*** The medical care market often generates positive externalities (spillover benefits). For example, an immunization against polio, smallpox, or measles benefits the immediate purchaser, but it also benefits society because it reduces the risk that other members of society will be infected with a highly contagious disease. Similarly, a healthy labor force is more productive, contributing to the general prosperity and well-being of society.
- ***Third-party payments: insurance*** Because four-fifths of all health care expenses are paid through public or private insurance, health care consumers pay much lower out-of-pocket "prices" than they would otherwise. Those lower prices are a distortion that results in "excess" consumption of health care services.

The Increasing Demand for Health Care

With these four features in mind, let's consider some factors that have increased the demand for health care over time.

Rising Incomes: The Role of Elasticities Because health care is a normal good, increases in domestic income have caused increases in the demand for health care. While there is some disagreement as to the exact income elasticity of demand for health care, several studies for industrially advanced countries suggest that the income elasticity coefficient is about 1. This means that per capita health care spending rises approximately in proportion to increases in per capita income. For example, a 3 percent increase in income will generate a 3 percent increase in health care expenditures. Some evidence suggests that income elasticity may be higher in the United States, perhaps as high as 1.5.

Estimates of the price elasticity of demand for health care imply that it is quite inelastic, with this coefficient being as low as 0.2. This means that the quantity of health care consumed declines relatively little as price increases. For example, a 10 percent increase in price would reduce quantity demanded by only 2 percent. An important consequence is that total health care spending will increase as the price of health care rises.

The relative insensitivity of health care spending to price changes results from four factors. First, people consider health care a necessity, not a luxury. Few, if any, good substitutes exist for medical care in treating injuries and infections and alleviating various ailments. Second, medical treatment is often provided in an emergency situation in which price considerations are secondary or irrelevant. Third, most consumers prefer a long-term relationship with their doctors and therefore do not "shop around" when health care prices rise. Fourth, most patients have insurance and are therefore not directly affected by the price of health care. If insured patients pay, for example, only 20 percent of their health care expenses, they are less concerned with price increases or price differences between hospitals and between doctors than they would be if they paid 100 percent themselves.

An Aging Population The U.S. population is aging. People 65 years of age and older constituted approximately 9 percent of the population in 1960 but 13.7 percent in 2012. Projections for the year 2030 indicate 20 percent of the population will be 65 or over by that year.

This aging of the population affects the demand for health care because older people encounter more frequent and more prolonged spells of illness. Specifically, those 65 and older consume about three and one-half times as much health care as those between 19 and 64. In turn, people over 84 consume almost two and one-half times as much health care as those in the 65 to 69 age group. Health care expenditures are often extraordinarily high in the last year of one's life.

In 2011 the oldest of the 76 million members of the baby boom generation born between 1946 and 1964 began turning 65. We can expect that fact to create a substantial surge in the demand for health care.

Unhealthy Lifestyles Substance abuse helps drive up health care costs. The abuse of alcohol, tobacco, and illicit drugs damages health and is therefore an important component of the demand for health care services. Alcohol is a major cause of injury-producing traffic accidents and liver disease. Tobacco use markedly increases the probability of cancer, heart disease, bronchitis, and emphysema. Illicit drugs are a major contributor to violent crime, health problems in infants, and the spread of AIDS. In addition, illicit-drug users make hundreds of thousands of costly visits to hospital emergency rooms each year. And overeating and lack of exercise contribute to heart disease, diabetes, and many other ailments. Obesity-related medical costs are estimated to be about $147 billion per year, with taxpayers picking up more than half the tab through Medicare and Medicaid.

The Role of Doctors Physicians may increase the demand for health care in several ways.

Supplier-Induced Demand As we mentioned before, doctors, the suppliers of medical services, have much more information about those services than consumers, who are the demanders. While a patient might be well informed about food products or more complex products such as cameras, he or she is not likely to be well informed about diagnostic tests such as magnetic resonance imaging or medical procedures such as joint replacements. Because of this asymmetric information (informational imbalance), a principal-agent problem emerges: The supplier, not the demander, decides what types and amounts of health care are to be consumed. This situation creates a possibility of "supplier-induced demand."

This possibility becomes especially relevant when doctors are paid on a **fee-for-service** basis, that is, paid separately for each service they perform. In light of the asymmetric information and fee-for-service arrangement, doctors have an opportunity and an incentive to suggest more health care services than are absolutely necessary (just as an auto repair shop has an opportunity and an incentive to recommend replacement of parts that are worn but still working).

More surgery is performed in the United States, where many doctors are paid a fee for each operation, than in foreign countries, where doctors are often paid fixed salaries unrelated to the number of operations they perform. Furthermore, doctors who own X-ray or ultrasound machines do four times as many tests as doctors who refer their patients to radiologists. More generally, studies suggest that up to one-third of common medical tests and procedures are either inappropriate or of questionable value.

The seller's control over consumption decisions has another result: It eliminates much of the power buyers

CONSIDER THIS . . .

Why Do Hospitals Sometimes Charge $25 for an Aspirin?

To save taxpayers money, Medicare and Medicaid set their payment rates for medical services above marginal cost but below average total cost. Doing so gives health care providers an incentive to provide services to Medicare and Medicaid patients because MR > MC. But it also means that government health insurance programs are not reimbursing the full cost of treating Medicare and Medicaid patients. In particular, the programs are not picking up their share of the fixed costs associated with providing health care.

As an example, consider an elderly person who uses Medicare. If he gets into a car accident and is taken to the local emergency room, the hospital will run up a wide variety of marginal costs, including ambulance charges, X-rays, medications, and the time of the nurses and doctors who help him. But the hospital also has a wide variety of fixed costs including rent, utility bills, computer networks, and lots of hideously expensive medical equipment.

These costs have to be borne by somebody. So when Medicare and Medicaid fail to pay their full share of the fixed costs, other patients must pick up the slack. The result has been for hospitals to transfer as much as possible of the fixed costs onto patients with private health insurance. The hospitals overbill private insurance companies so as to make up for the fixed costs that the government refuses to pay.

That is why you will hear stories about hospitals charging patients with private insurance $25 for a single aspirin or $100 for a newborn baby's first pair of diapers. They are making up for the fact that hospitals around the country lost nearly $28 billion in 2010 because Medicare and Medicaid on average only reimbursed hospitals about 92 percent of the total cost of providing medical services to Medicare and Medicaid patients.

might have in controlling the growth of health care prices and spending.

Defensive Medicine "Become a doctor and support a lawyer," says a bumper sticker. The number of medical malpractice lawsuits admittedly is high. To a medical doctor, each patient represents not only a person in need but also a possible malpractice suit. As a result, physicians tend to practice **defensive medicine.** They recommend more tests and procedures than are warranted medically or economically to protect themselves against malpractice suits.

Medical Ethics Medical ethics may drive up the demand for health care in two ways. First, doctors are legally and ethically committed to using "best-practice" techniques in serving their patients. This often means the use of costly medical procedures that may be of only slight benefit to patients.

Second, public values seem to support the medical ethic that human life should be sustained as long as possible. This makes it difficult to confront the notion that health care is provided with scarce resources and therefore must be rationed like any other good. Can society afford to provide $5,000-per-day intensive care to a comatose patient unlikely to be restored to reasonable health? Public priorities seem to indicate that such care should be provided, and those values again increase the demand for health care.

Role of Health Insurance

As we noted in Figure 22.1, 83 percent of health care spending is done not by health care consumers through direct out-of-pocket payments but by private health insurance companies or by the government through Medicare and Medicaid.

Individuals and families face potentially devastating monetary losses from a variety of hazards. Your house may burn down; you may be in an auto accident; or you may suffer a serious illness. An insurance program is a means of protection against the huge monetary losses that can result from such hazards. A number of people agree to pay certain amounts (premiums) periodically in return for the guarantee that they will be compensated if they should incur a particular misfortune. Insurance is a means of paying a relatively small known cost in exchange for obtaining protection against an uncertain but potentially much larger cost.

This financial arrangement can be highly advantageous to those purchasing insurance, but it also alters incentives in ways that can contribute to rising costs and the overconsumption of health care.

The Moral Hazard Problem The moral hazard problem is the tendency of one party to an agreement to alter her or his behavior in a way that is costly to the other party. Health care insurance can change behavior in two ways. First, some insured people may be less careful about their health, taking fewer steps to prevent accident or illness. Second, insured individuals have greater incentives to use health care more intensively than they would if they did not have insurance. Let's consider both aspects of moral hazard.

Less Prevention Health insurance may increase the demand for health care by encouraging behaviors that require more health care. Although most people with health care insurance are probably as careful about their health as are those without insurance, some may be more inclined to smoke, avoid exercise, and eat unhealthful foods, knowing they have insurance. Similarly, some individuals may take up ski jumping or rodeo bull riding if they have insurance covering the costs of orthopedic surgeons. And if their insurance covers rehabilitation programs, some people may be more inclined to experiment with alcohol or drugs.

Overconsumption Insured people go to doctors more often and request more diagnostic tests and more complex treatments than they would if they were uninsured because, with health insurance, the price or opportunity cost of consuming health care is minimal. For example, many individuals with private insurance pay a fixed premium for coverage, and beyond that, aside from a modest deductible, their health care is "free." This situation differs from most markets, in which the price to the consumer reflects the full opportunity cost of each unit of the good or service. In all markets, price provides a direct economic incentive to restrict use of the product. The minimal direct price to the insured consumer of health care, in contrast, creates an incentive to overuse the health care system. Of course, the penalty for overuse will ultimately show up in higher insurance premiums, but all policyholders will share those premiums. The cost increase for the individual health consumer will be relatively small.

Also, the availability of insurance removes the consumer's budget constraint (spending limitation) when he or she decides to consume health care. Recall from Chapter 7 that budget constraints limit the purchases of most products. But insured patients face minimal or no out-of-pocket expenditures at the time they purchase health care. Because affordability is not the issue, health care may be overconsumed.

Government Tax Subsidy Federal tax policy toward employer-financed health insurance works as a **tax subsidy** that strengthens the demand for health care services. Specifically, employees do not pay federal income or payroll tax (Social Security) on the value of the health insurance they receive as an employee benefit. Employees thus request and receive more of their total compensation as nontaxed health care benefits and less in taxed wages and salaries.

The government rationale for this tax treatment is that positive spillover benefits are associated with having a healthy, productive workforce. So it is appropriate to encourage health insurance for workers. The tax break does enable more of the population to have health insurance, but it also contributes to greater consumption of health care. Combined with other factors, the tax break may result in an overconsumption of health care.

To illustrate: If the marginal tax rate is, say, 28 percent, \$1 worth of health insurance is equivalent to 72 cents in after-tax pay. Because the worker can get more insurance for \$1 than for 72 cents, the exclusion of health insurance from taxation increases purchases of health insurance, thus increasing the demand for health care. In essence, the 28-cent difference acts as a government subsidy to health care. One estimate suggests that this subsidy costs the federal government \$120 billion per year in forgone tax revenue and boosts private health insurance spending by about one-third. Actual health care spending may be 10 to 20 percent higher than otherwise because of the subsidy.

Graphical Portrayal A simple demand and supply model illustrates the effect of health insurance on the health care market. Figure 22.3a depicts a competitive market for health care services; curve *D* shows the demand for health care services if all consumers are uninsured, and

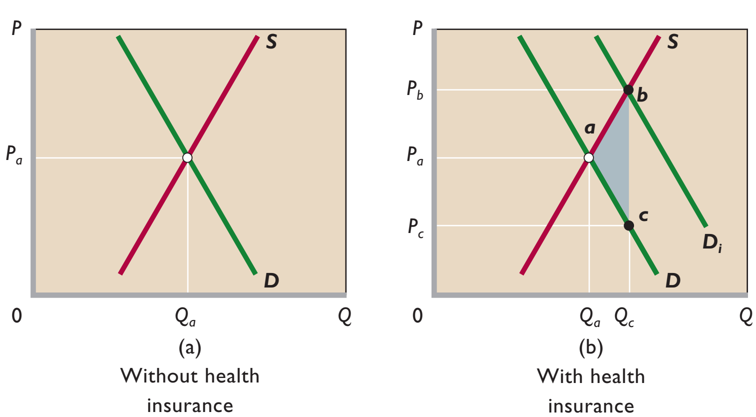


FIGURE 22.3 Insurance and the overallocation of resources to health care. (a) Without health insurance, the optimal amount of health care consumed is Q_a, where the marginal benefit and marginal cost of health care are equal. (b) The availability of private and public insurance increases the demand for health care, as from *D* to D_i, and reduces the price to the consumer from P_a to P_c (here, equal to one-third of the full price P_b). This lower after-insurance price results in overconsumption (Q_c rather than Q_a). Area *abc* represents the efficiency loss (or deadweight loss) from the overallocation of resources to health care.

S represents the supply of health care. At market price P_a the equilibrium quantity of health care is Q_a.

Recall from our discussion of competitive markets that output Q_a results in allocative efficiency, which means there is no better alternative use for the resources allocated to producing that level of health care. To see what we mean by "no better use," recall that:

- As we move down along demand curve D, each succeeding point indicates, via the price it represents, the marginal benefit that consumers receive from that unit.
- The supply curve is the producers' marginal-cost curve. As we move up along supply curve S, each succeeding point indicates the marginal cost of each successive unit of health care.
- For each unit produced up to the equilibrium quantity Q_a, marginal benefit exceeds marginal cost (because points on D are above those on S). At Q_a marginal benefit equals marginal cost, designating allocative efficiency. No matter what else those resources could have produced, the greatest net benefit to society is obtained by using those resources to produce Q_a units of health care.

But allocative efficiency occurs only when consumers pay the full market price for a product, as is assumed in Figure 22.3a. What happens when we introduce health insurance that covers, say, two-thirds of all health care costs? In Figure 22.3b, with private or public health insurance in place, consumers increase their demand for health care, as from D to D_i. At each possible price they desire more health care than before because insurance will pick up a large part of the bill. Given the supply curve of health care S, this increase in demand raises the price of health care to P_b. But with the insurance, consumers pay only one-third of the new higher price. This is less than without the insurance because the new price is only P_c ($= \frac{1}{3}P_b$)—rather than the previous price P_a. So they increase their consumption of health care from Q_a to Q_c.

The added consumption (and production) of health care is inefficient. Between Q_a and Q_c each unit's marginal cost to society (measured on curve S) exceeds its marginal benefit (measured on before-insurance demand curve D). Each unit of health care between Q_a and Q_c is an overallocation of resources to health care. Area *abc* shows the efficiency loss (or deadweight loss) that results.

Figure 22.3b implies that a trade-off exists between efficiency and equity. Standards of fairness or equity in the United States lead people to believe that all citizens should have access to basic health care, which is why government created social insurance in the form of Medicare and Medicaid. Also, it helps explain the federal tax subsidy to private health insurance, which again makes health care more accessible. The problem, as Figure 22.3b shows, is that the greater the availability of insurance (and thus the more equitable society makes access to health care), the greater the overallocation of resources to the health care industry. This overallocation would be even greater if health care were provided completely "free" under a program of national health insurance. Consumers would purchase health care as long as the marginal benefit to themselves as individuals was positive, regardless of the true cost to society.

Rationing to Control Costs We have just seen that by reducing the marginal costs facing patients, both private health insurance and government health insurance drive up prices. But then why is it that the United States, with its emphasis on private health insurance, spends so much more on health care than countries like the United Kingdom and Canada that provide national health insurance? If both types of insurance promote higher prices, why do we see higher health care spending in the United States than in those other countries?

One contributing factor is that the countries with national health insurance use various nonprice mechanisms to ration care. These mechanisms restrict the quantity of health care services supplied and, consequently, the amount of money spent providing health care.

Some nonprice rationing is done by committees of medical and budgetary experts. In the United Kingdom, for instance, the National Committee for Health and Clinical Excellence has set a general limit of £30,000 (approximately $44,000) on the cost of extending life for a year. Applying this rule to a specific situation, if an anticancer treatment would cost more than £30,000 to extend a cancer patient's life for a year, the United Kingdom's national health service would not pay for it. This cost rule keeps a lid on expenditures.

Waiting is another nonprice mechanism that rations care in countries with national health insurance. In the Canadian system, patients often have waits of weeks, months, or even years for certain diagnostic procedures and surgeries. This is the result of the Canadian government's effort to control expenditures by restricting hospitals' capital spending. To illustrate, there are only one-fifth as many magnetic resonance imaging (MRI) machines per million people in Canada as in the United States. This results in a substantial waiting list for MRI scans in Canada.

By contrast, private health insurers in the United States have not had to obey national committees that set spending limits. Nor have they had to answer to government budget officials attempting to control expenditures by restricting capital spending. Instead, private health insurers have faced a very different regulatory system that has tended to increase rather than decrease spending.

To see why this is true, note that until the Patient Protection and Affordable Care Act, private health insurers were regulated almost entirely at the state level. In addition, each state insurance regulator had very different incentives compared to the national health insurance regulators found in other countries.

In a country with national health insurance, regulators are confronted by their government's budget constraint and the fact that the government has a limited supply of tax dollars to spend on health care. As a result, they have a strong incentive to deny care and limit spending. By contrast, state insurance regulators in the United States did not have to worry about exhausting government budgets because private health insurance is paid for with private money rather than government money. This led to a tendency for state insurance regulators to focus on benefits more than costs.

To see why this happened, note that requiring insurance companies to cover more people or treat more conditions is politically popular, while requiring them to cut coverage to save money is politically unpopular. With many states having popularly elected state insurance commissioners and the rest having insurance commissioners appointed by either the governor or the legislature, there was constant political pressure for state insurance regulators to pass new regulations requiring insurance companies to spend more rather than less. States incrementally imposed various rules expanding the number of conditions that had to be covered by insurers as well as the amounts that insurers had to spend on patient care.

Because these requirements involved the costly provision of additional care, insurance companies responded by raising insurance premiums. Thus, America's state-based system of government insurance regulation tended to increase rather than decrease the amount spent on health care.

Many economists view this regulatory system as one of the factors that has contributed to the United States spending more of its GDP on health care than any other nation on earth. While regulators in other countries have sought out ways to deny care and reduce costs, state regulators in America have tended to mandate that insurance companies expand treatment and incur additional costs.

As we will discuss later, the Patient Protection and Affordable Care Act creates a new set of federal insurance regulators that will largely supersede state insurance regulators. Part of the controversy related to the bill is whether the new federal regulators might eventually be tasked with denying care in the way that European regulators have.

Supply Factors in Rising Health Care Prices

Supply factors have also played a role in rising health care prices. Specifically, the supply of health care services has increased, but more slowly than demand. A combination of factors has produced this relatively slow growth of supply.

Supply of Physicians The supply of physicians in the United States has increased over the years; in 1975 there were 169 physicians per 100,000 people; by 2011 there were 275. But this increase in supply has not kept up with the increase in the demand for physicians' services. As a result, physicians' fees and incomes have increased more rapidly than average prices and incomes for the economy as a whole.

Conventional wisdom has been that physician groups, for example, the American Medical Association, have purposely kept admissions to medical schools, and therefore the supply of doctors, artificially low. But that is too simplistic. A rapidly rising cost of medical education seems to be the main cause of the relatively slow growth of doctor supply. Medical training requires 4 years of college, 4 years of medical school, an internship, and perhaps 3 or 4 years of training in a medical specialty. The opportunity cost of this education has increased because the salaries of similarly capable people have soared in other professions. The direct expenses have also increased, largely due to the increasingly sophisticated levels of medical care and therefore of medical training.

High and rising education and training costs have necessitated high and rising doctors' fees to ensure an adequate return on this sort of investment in human capital. Physicians' incomes are indeed high, averaging in 2011 from about $178,000 for family care physicians up to $541,000 for neurosurgeons. But the costs of obtaining the skills necessary to become a physician are also very high. Data show that while doctors have high rates of return on their educational expenses, those returns are below the returns for the holders of masters of business administration degrees.

Slow Productivity Growth Productivity growth in an industry tends to reduce costs and increase supply. In the health care industry, such productivity growth has been modest. One reason is that health care is a service, and it is generally more difficult to increase productivity for services than for goods. It is relatively easy to increase productivity in manufacturing by mechanizing the production process. With more and better machinery, the same number of workers can produce greater output. But services often are a different matter. It is not easy, for example, to mechanize haircuts, child care, and pizza delivery.

How do you significantly increase the productivity of physicians, nurses, and home care providers?

Also, competition for patients among many providers of health care has not been sufficiently brisk to force them to look for ways to reduce costs by increasing productivity. When buying most goods, customers typically shop around for the lowest price. This shopping requires that sellers keep their prices low and look to productivity increases to maintain or expand their profits. But patients rarely shop for the lowest prices when seeking medical care. In fact, a patient may feel uncomfortable about being operated on by a physician who charges the lowest price. Moreover, if insurance pays for the surgery, there is no reason to consider price at all. The point is that unusual features of the market for health care limit competitive pricing and thus reduce incentives to achieve cost saving via advances in productivity.

Changes in Medical Technology Some technological advances in medicine have lowered costs. For example, the development of vaccines for polio, smallpox, and measles has greatly reduced health care expenditures for the treatment of those diseases. And reduced lengths of stays in hospitals have lowered the costs of medical care.

But many medical technologies developed since the Second World War have significantly increased the cost of medical care either by increasing prices or by extending procedures to a greater number of people. For example, because they give more accurate information, advanced body scanners costing up to $1,000 per scan are often used in place of X-rays that cost less than $100 for each scan. Desiring to offer the highest quality of service, hospitals want to use the very latest equipment and procedures. These newer, more expensive treatments are believed to be more effective than older ones. But doctors and hospital administrators both realize that the high fixed cost of such equipment means it must be used extensively to reduce the average cost per patient and recoup the investment at "fair return" charges per procedure.

As another example, organ transplants are extremely costly. Before the development of this technology, a person with a serious liver malfunction died. Now a liver transplant can cost $200,000 or more, with subsequent medical attention costing $10,000 to $20,000 per year for the rest of the patient's life.

Finally, consider new prescription medications. Pharmaceutical companies have developed very expensive drugs that often replace less expensive ones and are prescribed for a much wider range of physical and mental illnesses. Although these remarkable new medications greatly improve health care, they also contribute to rising health care costs.

CONSIDER THIS . . .

Electronic Medical Records

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 provides $20 billion of subsidies to encourage hospitals and physicians to adopt electronic medical records.

The subsidies are designed to get doctors and nurses to switch from traditional paper records to electronic databases. That way, hospital administrators should be able to easily search for instances where costs might be cut. They could, for instance, notice whether doctors are overprescribing certain medications or scheduling expensive MRI scans when less-expensive X-rays would suffice.

Unfortunately, these systems seem to raise costs and lower the quality of care. When records are done with pen and paper, doctors only write down the information they think is relevant. By contrast, electronic systems demand that during every appointment, the doctor fill out every single item from a comprehensive list of questions and drop-down menus. That is wasteful because the vast majority of those items are going to be totally irrelevant for any particular patient. The systems are so time-intensive that many doctors report that they are now seeing fewer patients because it takes so long to fill out the forms.

At the same time, insurance billings have gone up because hospitals using electronic systems now have a comprehensive record of *everything* that was discussed at each appointment. Armed with that information, they are billing insurance companies and Medicare and Medicaid for many more procedures than was the case when records were kept by hand and doctors only wrote down the most important information about each patient.

The projected efficiency gains from being able to search for patterns of overuse have not materialized because the financial incentive for each hospital is not to look for patterns of overuse but, instead, to bill for every little thing that it can. As a result, care is down while costs are up.

The historical willingness of private and public insurance to pay for new treatments without regard to price and number of patients has contributed to the incentive to develop and use new technologies. Insurers, in effect, have encouraged research into and development of health care technologies, regardless of their cost. Recently, when insurance companies resisted paying for new expensive treatments such as bone marrow transplants, public outcries led them to change their minds. So expanding insurance coverage leads to new, often more expensive medical technologies, which in turn lead to a demand for a wider definition of what should be covered by insurance.

Relative Importance

According to most analysts, the demand and supply factors we have discussed vary in their impact on escalating health care costs. As we noted, the income elasticity of demand for health care is estimated to be upward of 1.5 in the United States, meaning that increased income brings with it more-than-proportionate increases in health care spending. But rising income does not alone explain the rocketing increase in health care spending as a percentage of total domestic output (income). Furthermore, government studies estimate that the aging population accounts for less than 10 percent of the current increase in per capita health care spending.

Most experts attribute the relative rise in health care spending to (1) advances in medical technology, combined with (2) the medical ethic of providing the best treatment available, (3) private and public health insurance, and (4) fee-for-service physicians' payments by health insurance firms. Through technological progress, great strides have been made in the diagnosis, treatment, and prevention of illness. But the third-party (insurance) payment system provides little incentive to limit the development or use of such technologies because it has no mechanism to force an equating of marginal costs and marginal benefits. And the "best treatment available" ethic, together with the fee-for-service payment system, ensures that any new technology with a positive marginal benefit will get used and be billed for, regardless of the marginal cost to society.

QUICK REVIEW 22.2

- Characteristics of the health care market are (a) the widespread view of health care as a "right," (b) asymmetric information between consumers and suppliers, (c) the presence of positive externalities, and (d) payment mostly by insurance.
- The demand for health care has increased for many reasons, including rising incomes, an aging population, unhealthy lifestyles, the role of doctors as advisers to patients, the practice of defensive medicine, and a fee-for-service payment system via health insurance.
- Countries with national health insurance systems contain costs by denying care for certain procedures and by limiting capital expenditures.
- The supply of health care has grown slowly, primarily because of (a) relatively slow productivity growth in the health care industry, (b) rising costs of medical education and training, and (c) greater use of very-high-cost health care technologies.

Cost Containment: Altering Incentives

LO22.5 Describe the cost-containment strategies that rely on altering the financial incentives facing either patients or health service providers.

The Patient Protection and Affordable Care Act is the latest in a long series of attempts to control the growth of health care costs, prices, and spending. Many of these efforts have been focused on reducing the incentives to overconsume health care.

Deductibles and Copayments

Insurance companies have reacted to rising health care costs by imposing sizable deductibles and copayments on those they insure. Instead of covering all of an insured's medical costs, a policy might now specify that the insured pay the first $250 or $500 of each year's health care costs (the deductible) and 15 or 20 percent of all additional costs (the copayment). The deductible and copayment are intended to alleviate the overuse problem by creating a direct payment and therefore an opportunity cost to the health care consumer. The deductible has the added advantage of reducing the administrative costs of insurance companies in processing many small claims.

Health Savings Accounts

A federal law enacted in 2003 established **health savings accounts (HSAs).** These accounts are available to all workers who are covered by health insurance plans with annual deductibles of $1,000 or more and do not have other first-dollar insurance coverage. Individuals can make tax-deductible contributions into their HSAs, even if they do not itemize deductions on their tax forms. Employers can also make tax-free contributions to workers' accounts if they choose. Earnings on the funds in HSAs are not taxable, and the owners of these accounts can use them to pay for approved medical expenses. Unused funds in HSAs accumulate and remain available for later out-of-pocket medical expenses. Account holders can place additional money into their accounts each year between age 55 and the year they become eligible for Medicare.

HSAs are designed to promote personal saving out of which workers can pay routine health care expenses while working and Medicare copayments and deductibles later, during retirement. HSAs are also designed to reduce escalating medical expenses by injecting an element of competition into health care delivery. Because individuals are using some of their own HSA money to pay for health care, they presumably will assess their personal marginal costs and marginal benefits in choosing how much and what type of health care to obtain. They will also have a strong incentive to inquire about and compare prices charged by various qualified medical providers. Holders of HSAs never lose their accumulated funds. They can remove money for nonmedical purchases but must pay income taxes and a 10 percent penalty on such withdrawals.

Managed Care

Managed-care organizations (or systems) are those in which medical services are controlled or coordinated by insurance companies or health care organizations to reduce health care expenditures. In 2009 nearly 90 percent of all U.S. workers received health care through such "managed care." These organizations are of two main types.

Some insurance companies have set up **preferred provider organizations (PPOs),** which require that hospitals and physicians accept discounted prices for their services as a condition for being included in the insurance plan. The policyholder receives a list of participating hospitals and doctors and is given, say, 80 to 100 percent reimbursement of health care costs when treated by PPO physicians and hospitals. If a patient chooses a doctor or hospital outside the PPO, the insurance company reimburses only 60 to 70 percent. In return for being included as a PPO provider, doctors and hospitals agree to rates set by the insurance company for each service. Because these fees are less than those usually charged, PPOs reduce health insurance premiums and health care expenditures.

Many Americans now receive their medical care from **health maintenance organizations (HMOs),** which provide health care services to a specific group of enrollees in exchange for a set annual fee per enrollee. HMOs employ their own physicians and contract for specialized services with outside providers and hospitals. They then contract with firms or government units to provide medical care for their workers, who thereby become HMO members. Because HMOs have fixed annual revenue, they may lose money if they provide "too much" care. So they have an incentive to hold down costs. They also have an incentive to provide preventive care in order to reduce the potentially far larger expense of corrective care.

Both PPOs and HMOs are managed-care organizations because medical use and spending are "managed" by closely monitoring physicians' and hospitals' behavior. The purpose of close monitoring is to eliminate unnecessary tests and treatments. Doctors in managed-care organizations might not order an MRI scan or an ultrasound test or suggest surgery because their work is monitored and because they may have a fixed budget. In contrast, an independent fee-for-service physician facing little or no oversight may have a financial incentive to order the test or do the surgery. Doctors and hospitals in a managed-care organization often share in an "incentive pool" of funds when they meet their cost-control goals.

The advantages of managed-care plans are that they provide health care at lower prices than traditional insurance and emphasize preventive medicine. The disadvantages are that the patient usually is restricted to physicians employed by or under contract with the managed-care plan. Also, some say that the focus on reducing costs has gone too far, resulting in denial of highly expensive, but effective, treatment. This "too far" criticism was mainly leveled at HMOs, where incentives to reduce costs were the greatest. Perhaps because of a backlash against HMOs, firms have increasingly shifted workers from managed care toward PPOs.

Medicare and DRG

In 1983 the federal government altered the way it makes payments for hospital services received by Medicare patients. Rather than automatically paying all costs related to a patient's treatment and length of hospital stay, Medicare authorized payments based on a **diagnosis-related-group (DRG) system.** Under DRG, a hospital receives a fixed payment for treating each patient; that payment is an amount associated with the diagnosis—one of several hundred carefully detailed diagnostic categories—that best characterizes the patient's condition and needs.

DRG payments obviously give hospitals the incentive to restrict the amount of resources used in treating patients. It is no surprise that under DRG the length of hospital stays has fallen sharply and more patients are treated on an outpatient basis. Critics, however, argue that this is evidence of diminished health care quality.

Limits on Malpractice Awards

Congress has, for many years, debated whether to cap (at, say, $250,000 or $500,000) the "pain and suffering" awards on medical malpractice lawsuits against physicians. Those who support malpractice caps say that patients should receive full compensation for economic losses but not be made wealthy through huge jury awards. They contend that capping the awards will reduce medical malpractice premiums and therefore lower health care costs. Opponents of caps counter that large "pain and suffering" awards deter medical malpractice. If so, such awards improve the overall quality of the health care system. Opponents also point out that malpractice awards are a negligible percentage of total health care costs. Thirty-three states have placed caps on the "pain and suffering" portion of malpractice awards.

QUICK REVIEW 22.3

- Policy makers have pursued several strategies when attempting to contain health care spending and prices.
- Insurance deductibles and copayments confront consumers with opportunity costs; managed-care organizations attempt to restrict their members' use of health services; the diagnosis-related-group (DRG) system caps the amount Medicare will spend on any procedure; and health savings accounts (HSAs) confront individuals with the marginal cost of routine health care expenses.

The Patient Protection and Affordable Care Act

LO22.6 Summarize the goals of the Patient Protection and Affordable Care Act and the major changes that it institutes.

The primary goal of the Patient Protection and Affordable Care Act (PPACA) passed in 2010 was not cost containment but rather the extension of health insurance coverage to all Americans. In truth, covering every single citizen would have been possible only if America had moved to a national health insurance system similar to that used in Canada. Such a move would have been impossible politically, however, because opinion polls indicated that 75 percent of Americans with employer-provided health insurance rated their coverage as good or very good.

Thus, President Obama and like-minded members of Congress did not pursue the creation of a national health insurance system. Rather, they moved to extend and expand the existing system in which nearly all Americans receive their health care through either employer-provided health insurance or government-provided health insurance (Medicaid and Medicare). In promoting the PPACA, the president reassured audiences by telling them, "If you like your health care plan, you can keep it."

Major Provisions

The authors of the PPACA understood that extending insurance coverage to millions of previously uninsured people would be costly. As with any group of people enrolled in a health insurance program, many would eventually become sick and need costly treatments. This problem was exacerbated by the fact that many of those without insurance were known to suffer from extremely costly medical conditions. Indeed, these individuals were without insurance precisely because private insurance companies (which have to either break even or go bankrupt) considered them too expensive to insure. Thus, if those with costly medical conditions were to be covered, significant new revenue sources had to be found.

The PPACA aimed to obtain the needed revenue from two main sources: a personal mandate to buy insurance and an assortment of other new taxes. We will discuss each as we go over the PPACA's major provisions.

Preexisting Conditions, Caps, and Drops The PPACA makes it illegal for insurance companies to deny coverage to anyone on the basis of a preexisting medical condition. As just discussed, this ban will lead to the enrollment of millions of individuals with costly health conditions, as these individuals will be almost certain to enroll as soon as the ban goes into effect.

The PPACA also increases the amount of money that insurance companies will be spending in the future by prohibiting them from imposing annual or lifetime expenditure caps.

To prevent insurance companies from dropping policyholders just because they develop a costly illness, the PPACA also makes fraud the only legal reason that an insurance company can drop a policyholder. Because this provision will force insurance companies to keep very sick people enrolled, it too is expected to entail significant cost increases for insurance companies.

Employer Mandate The PPACA has an **employer mandate** (requirement) that every firm with 50 or more full-time employees must either purchase health insurance for each of their full-time employees or pay a fine of $2,000 per employee. This provision of the law is intended to extend employer-paid health insurance to as many workers as possible so that private employers, rather than the government, will bear as much of the cost of extending insurance coverage to the uninsured as possible.

Firms with fewer than 50 full-time employees are exempt from the employer mandate because the high cost of health insurance might bankrupt many smaller firms.

Personal Mandate The PPACA contains a **personal mandate** (requirement) that individuals must purchase health insurance for themselves and their dependents unless they are already covered by either government insurance or employer-provided insurance. Anyone refusing will be fined. The fine is the larger of either $695 per uninsured family member or 2.5 percent of family income.

As we will explain next, the PPACA contains extensive subsidies to ensure that poorer people will not be financially devastated by the personal mandate's requirement to buy health insurance. However, it must be understood that the point of the personal mandate is to force higher-income healthy people (especially healthy young workers) to buy health insurance so that their insurance premiums can help pay for the high health care bills of those with costly medical conditions as well as the subsidies needed to make health insurance affordable for those with lower incomes.

Covering the Poor Millions of poorer Americans lacked private health insurance either (*a*) because they were unemployed and thus not receiving employer-provided health insurance or (*b*) because their employers did not provide health insurance. Some of these poorer Americans could obtain government health insurance through either Medicaid or Medicare, but those who did not found themselves without any health insurance, either private or public.

The PPACA attempts to cover those with lower incomes in three ways. First, the employer mandate will induce many larger employers to provide insurance for all of their full-time employees, including the poorer ones. Second, the law expands the Medicaid system to cover anyone whose income is less than 133 percent of the poverty level. Third, the PPACA subsidizes the purchase price of health insurance for those who must buy their own health insurance to comply with the individual mandate.

The subsidies actually extend well into the middle class because they extend upward along the income scale even to persons making up to four times the federal poverty level. Taking into account the fact that individuals and households have different federal poverty levels, the subsidies would extend to individuals making up to about $44,000 per year and families of four making up to about $88,000 per year. By comparison, full-time workers had median annual earnings of about $41,560 in 2011, while the median family income in 2011 was about $50,054. The subsidies extend into the upper half of the income distribution because health insurance is so expensive that the personal mandate would have been financially ruinous for even middle-income workers if it had not been accompanied by subsidies.

A complicated formula adjusts the size of the subsidies by income level. The formula kicks in at 133 percent of the federal poverty level because anyone earning less will receive free government health insurance through Medicaid. For those earning slightly more than 133 percent of the poverty level (about $15,000 for an individual and about $30,700 for a family of four), the subsidies would be large enough so that they would not have to spend more than about 4 percent of their incomes purchasing health insurance. The subsidies get progressively less generous as incomes rise, so that those earning three to four times the poverty level would be subsidized such that they would have to pay about 10 percent of their incomes to buy health insurance.

CONSIDER THIS . . .

PPACA Implementation Problems

As this book was going to press, the PPACA was in the initial stages of being implemented. Several problems were apparent, including:

- Many states had declined to set up their own insurance exchanges, thereby leaving the task up to the federal government.
- The haste with which the law was written had led to errors, including an omission that disqualified many poor workers from receiving subsidies to help them deal with the high cost of the personal mandate.
- Many employers were reacting to the very expensive employer mandate by limiting some or all of their employees to working less than 30 hours per week.

Employers reacted that way because the employer mandate (to either buy employees health insurance or pay a fine) only applies to "full-time employees," which the PPACA defines as employees who work more than 30 hours per week. Thus, by downsizing workers to less than 30 hours per week, employers could avoid the employer mandate altogether.

Firms had a strong financial incentive to do so because the average cost of a family insurance policy in 2012 was $15,700. For many firms, buying such a policy for each full-time worker would have meant bankruptcy. Hence their desire to avoid the employer mandate by cutting employee hours.

As a result of so many firms downsizing full-time positions, many people who wanted to work more than 30 hours per week were finding that the only way to do so was by stringing together multiple part-time jobs. In addition, they were finding that as far as their personal situations were concerned, the main intention of the PPACA wasn't being accomplished. They still didn't have health insurance!

Insurance Exchanges Individuals shopping for their own insurance will do so in government-regulated markets called **insurance exchanges.** There will be one exchange for each state, and federal regulators will only allow policies meeting certain standards to be offered. The regulators cannot set prices directly but will have the authority to withdraw approval from any insurer that requests a price increase that regulators deem to be unjustified on the basis of higher costs. It is hoped that the exchanges will reduce the growth of health care spending by fostering competition among insurance companies.

Other Provisions The 2,400-page PPACA contains hundreds of additional provisions. Some of the more publicized include:

- Mandating that the adult children of parents with employer-provided health insurance remain covered by their parents' insurance through age 26.
- Making it illegal for insurance companies to charge copayments or apply deductibles to annual check ups or preventive care.
- Requiring insurers to spend at least 80 percent of the money they receive in premiums on either health care or improving health care.

Taxes To help pay for the extension of health insurance to millions of previously uninsured people, the PPACA imposes several new taxes. The more prominent are:

- A 0.9 percentage point increase in the Medicare payroll tax for individuals earning more than $200,000 per year and for married couples earning more than $250,000 per year.
- A 3.8 percentage point increase in the capital gains tax for individuals earning more than $200,000 per year and for married couples earning more than $250,000 per year.
- A 40 percent tax payable by employers on any employer-provided insurance policy whose premium exceeds $10,200 per year for individual coverage or $27,500 per year for family coverage.
- A 2.9 percent excise tax applied to everything sold by medical device manufacturers.
- A 10 percent tax levied on indoor tanning.

Objections and Alternatives

The PPACA was strongly opposed and passed Congress without a single approving vote in either chamber of Congress from members of the minority (Republican) party. The legislation also failed to achieve majority support in public opinion polls conducted on the eve of the legislation's passage.

Some of those voicing objections worried that federal control over the pricing and content of insurance policies would lead to greater inefficiencies in health care by adding additional layers of bureaucracy. Others objected because they felt that the PPACA might be the first step toward the creation of a national health insurance system in which nonprice rationing might become necessary to hold down expenditures. Yet others pointed to financial projections indicating that the revenue sources legislated by the PPACA would not be nearly sufficient to cover future health care expenses, especially over the longer run.

An additional concern was whether the PPACA would reduce the growth rate of health care expenditures and thereby fulfill the president's promise that the law would "bend the cost curve down." Many economists worried that the large subsidies provided by the law would raise prices and increase consumption (as in Figure 22.3 and the nearby discussion). With even middle-class individuals and families eligible for significant government subsidies, inefficient health care spending might increase significantly.

As an alternative, some opponents of the PPACA pointed to the health care system in Singapore and recent experiments with the health insurance offered to employees of the State of Indiana. Both systems reduce wasteful expenditures by increasing the percentage of health care spending that comes directly out of consumer's pockets, thereby forcing them to consider opportunity costs and weigh marginal benefits against marginal costs. (See this chapter's Last Word for more.)

In evaluating the pros and cons of the PPACA, one thing seems clear. It will not be the last word on health care reform in America. Indeed, the economic challenges related to health care will only get stronger. The combination of an aging population and advances in medical technology seem to be on a collision course with the reality of economic scarcity such that individuals and society will face increasingly difficult choices about how much health care to consume and how to pay for it.

QUICK REVIEW 22.4

- The Patient Protection and Affordable Care Act (PPACA) is an attempt to extend either private or public insurance coverage to all U.S. citizens and legal residents.
- The PPACA includes (1) a *personal mandate* that requires all citizens and legal residents to purchase insurance coverage for themselves and their dependents if they are not already provided with insurance by their employer or by the government as well as (2) an *employer mandate* that requires all firms with more than 50 full-time employees to either offer health insurance coverage to their full-time employees or pay large fines.
- The PPACA also bans insurance companies from denying coverage on the basis of preexisting conditions; includes various subsidies so that the personal mandate will not bankrupt the poor and middle class; provides for the creation of state insurance exchanges where individuals can comparison shop for government-approved health insurance policies; and imposes various taxes to help pay for the increased expenditures that will be required to extend insurance coverage to the previously uninsured.

LAST WORD

Singapore's Efficient and Effective Health Care System

How Does Singapore Deliver Some of the Best Health Care in the World While Spending Nearly 80 Percent Less per Person Than the United States?

In every health-quality category monitored by the World Health Organization, the small island nation of Singapore is either number one in the world or near the top of the list. Among other achievements, Singapore has the world's lowest rate of infant mortality and the world's fourth highest life expectancy.

One might expect that achieving these exceptional outcomes would be extremely expensive. But Singapore is also number one in another category. It spends less per person on health care than any other developed nation. In 2011 the United States spent 17.9 percent of its GDP on health care. Singapore spent just 4.0 percent.

How does Singapore deliver world-class health care while spending less than any other developed nation? The answer is a unique combination of government mandates to encourage competition, high out-of-pocket costs for consumers, and laws requiring people to save for future health expenditures.

Competition is encouraged by forcing hospitals to post prices for each of their services. Armed with this information, patients can shop around for the best deal. The government also publishes the track record of each hospital on each service so that consumers can make informed decisions about quality as well as price. With consumers choosing on the basis of cost and quality, local hospitals compete to reduce costs and improve quality.

Singapore also insists upon high out-of-pocket costs to avoid the overconsumption and high prices that result when insurance policies pick up most of the price for medical procedures. Indeed, out-of-pocket spending represents about 92 percent of all non-government health care spending in Singapore, compared to just 11 percent in the United States.

Having to pay for most medical spending out of pocket, however, means that Singapore's citizens are faced with having to pay for most of their health care themselves. How can this be done without bankrupting the average citizen? The answer is mandatory health savings accounts.

Singapore's citizens are required to save about 6 percent of their incomes into "MediSave" accounts. MediSave deposits are private property, so people have an incentive to spend the money in their accounts wisely. In addition, the citizens of Singapore also know that they won't be left helpless if the money in their MediSave accounts runs out. The government subsidizes the

SUMMARY

LO22.1 Convey important facts about rising health care costs in the United States.

The U.S. health care industry comprises 17 million workers (including about 850,000 practicing physicians) and 5,800 hospitals.

Unlike nations with publicly funded systems of national health insurance, the United States delivers a large fraction of its health care through private, employer-provided health insurance.

U.S. health care spending has increased both absolutely and as a percentage of GDP.

LO22.2 Relate the economic implications of rising health care costs.

Rising health care costs and prices have (*a*) reduced access to the health care system, (*b*) contributed to slower real wage growth and expanded the employment of part-time and temporary workers, and (*c*) caused governments to restrict spending on non-health programs and to raise taxes.

The core of the health care problem is an alleged overallocation of resources to the health care industry.

LO22.3 Discuss the problem of limited access to health care for those without insurance.

About 49 million Americans, or 16 percent of the population, did not have health insurance in 2011. The uninsured were concentrated among the poor, the chronically ill, the unemployed, the young, those employed by small firms, and low-wage workers.

LO22.4 List the demand and supply factors explaining rising health care costs.

Special characteristics of the health care market include (*a*) the belief that health care is a "right," (*b*) an imbalance of information between consumers and suppliers, (*c*) the presence of positive externalities, and (*d*) the payment of most health care expenses by private or public insurance.

health care of those who have exhausted their MediSave accounts as well as the health care of the poor and others who have not been able to accumulate much money in their MediSave accounts.

Could elements of Singapore's system help to hold down medical costs in the United States? Two cases suggest that the answer is yes.

First, consider the health care plan offered by Whole Foods Markets to its employees. The company deposits $1,800 per year into a "personal wellness account" for each of its full-time employees. It simultaneously pays for a high-deductible health insurance plan that will pick up 100 percent of all medical expenses exceeding $2,500 in a given year. This combination implies that employees are *at most* on the hook for $700 a year—that is, for the difference between the $1,800 in their personal wellness accounts and the $2,500 deductible on their health insurance policy (above which, all medical expenses are covered).

Because both the money in the personal wellness account as well as the $700 that employees might have to spend before reaching the $2,500 deductible are personal property, Whole Foods Markets has effectively created a system in which all medical spending up to $2,500 is an out-of-pocket expense. This forces employees to examine the opportunity cost of any potential medical expenditure. The result is less spending.

A similar plan offered to employees of the State of Indiana puts $2,750 per year into a health savings account and then provides an insurance policy that covers 80 percent of any medical expenses between $2,750 and $8,000 and 100 percent of any expenses above $8,000. The Indiana plan's design means that any state employee volunteering for the plan must pay 100 percent of all spending up to $2,750 from their health savings accounts. As with Singapore's system and Whole Foods' system, this encourages prudence. The result has been a 35 percent decline in total health care spending for those who volunteered for the plan versus state employees who opted to stick with the state's traditional PPO option. In addition, an independent audit showed that participants in the new plan were not cutting corners by skimping on preventive care like annual mammogram screenings for cancer. Thus, the savings appear to be permanent and sustainable.

The program is also popular, with positive personal recommendations causing voluntary participation to rise from 2 percent of state employees in the program's first year to 70 percent of state employees in the program's second year.

While rising incomes, an aging population, and substance abuse have all contributed to an increasing demand for health care, the role of doctors is also significant. Because of asymmetric information, physicians influence the demand for their own services. The fee-for-service payment system, combined with defensive medicine to protect against malpractice suits, also increases the demand for health care.

The moral hazard problem arising from health insurance takes two forms: (*a*) people may be less careful of their health and (*b*) there is an incentive to overconsume health care.

The exemption of employer-paid health insurance from the federal income tax subsidizes health care. The subsidy increases demand, leading to higher prices and a likely overallocation of resources to health care.

Countries with systems of national health insurance also increase demand by subsidizing health care. Facing limited budgets, those countries engage in nonprice rationing to restrict health care expenditures. Rationing mechanisms include waiting lists, committees that set standards for denial of service, and restrictions on capital spending.

Because private insurance does not involve government expenditures, the state regulators charged with regulating private insurance companies in the United States focus more on expanding politically popular benefits than on restricting costs.

Slow productivity growth in the health care industry and, more important, cost-increasing advances in health care technology have restricted the expansion of the supply of medical care and have boosted prices.

LO22.5 Describe the cost-containment strategies that rely on altering the financial incentives facing either patients or health service providers.

Strategies that have attempted to contain health care prices and spending include (*a*) insurance deductibles and copayments to confront consumers with opportunity costs, (*b*) managed-care organizations—preferred provider organizations (PPOs) and health maintenance organizations (HMOs)—that attempt to restrict their members' use of health services, (*c*) the diagnosis-related-group (DRG) system that caps the amount Medicare will spend on any procedure, and (*d*) health savings accounts (HSAs) that also confront individuals with opportunity costs when they spend out of their tax-free HSA accounts.

LO22.6 Summarize the goals of the Patient Protection and Affordable Care Act and the major changes that it institutes.

The Patient Protection and Affordable Care Act (PPACA) of 2010 is an attempt to extend either private or public (Medicare and Medicaid) insurance coverage to all U.S. citizens and legal residents.

Enrolling millions of previously uninsured people (including the chronically ill) into health insurance will be costly, so the PPACA includes a personal mandate that requires all citizens and legal residents to purchase insurance coverage for themselves and their dependents if they are not already provided with insurance by their employer or by the government. The goal is to compel healthy people to purchase insurance so that their premiums can help to pay for the health care costs of the previously uninsured (many of whom are likely to be chronically ill.)

The PPACA also (*a*) bans insurance companies from denying coverage on the basis of preexisting conditions, (*b*) includes various subsidies so that the personal mandate will not bankrupt the poor and middle class, (*c*) provides for the creation of state insurance exchanges where individuals can comparison shop for government-approved health insurance policies, and (*d*) imposes various taxes to help pay for the increased expenditures that will be required to extend insurance coverage to the previously uninsured.

TERMS AND CONCEPTS

Patient Protection and Affordable Care Act (PPACA)
national health insurance
deductibles
copayments
fee for service
defensive medicine
tax subsidy
health savings accounts (HSAs)
preferred provider organizations (PPOs)
health maintenance organizations (HMOs)
diagnosis-related-group (DRG) system
employer mandate
personal mandate
insurance exchanges

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Why would increased spending as a percentage of GDP on, say, household appliances or education in a particular economy be regarded as economically desirable? Why, then, is there so much concern about rising expenditures as a percentage of GDP on health care? **LO22.1**
2. What are the "twin problems" of the health care industry as viewed by society? How are they related? **LO22.1**
3. Briefly describe the main features of Medicare and Medicaid, indicating how each is financed. **LO22.1**
4. What are the implications of rapidly rising health care prices and spending for (*a*) the growth of real wage rates, (*b*) government budgets, and (*c*) offshoring of U.S. jobs? Explain. **LO22.2**
5. What are the main groups without health insurance? **LO22.3**
6. List the special characteristics of the U.S. health care market and specify how each affects health care problems. **LO22.3**
7. What are the estimated income and price elasticities of demand for health care? How does each relate to rising health care costs? **LO22.4**
8. Briefly discuss the demand and supply factors that contribute to rising health costs. Specify how (*a*) asymmetric information, (*b*) fee-for-service payments, (*c*) defensive medicine, and (*d*) medical ethics might cause health care costs to rise. **LO22.4**
9. How do advances in medical technology and health insurance interact to drive up the cost of medical care? **LO22.4**
10. Using the concepts in Chapter 7's discussion of consumer behavior, explain how health care insurance results in an overallocation of resources to the health care industry. Use a demand and supply diagram to specify the resulting efficiency loss. **LO22.4**
11. How is the moral hazard problem relevant to the health care market? **LO22.4**
12. What is the rationale for exempting a firm's contribution to its workers' health insurance from taxation as worker income? What is the impact of this exemption on allocative efficiency in the health care industry? **LO22.5**
13. What are (*a*) preferred provider organizations and (*b*) health maintenance organizations? In your answer, explain how each is designed to alleviate the overconsumption of health care. **LO22.5**
14. What are health savings accounts (HSAs)? How might they reduce the overconsumption of health care resulting from traditional insurance? How might they introduce an

element of price competition into the health care system? **LO22.5**

15. Why is the PPACA's attempt to extend insurance coverage to all Americans so costly? How does the PPACA attempt to obtain the funds needed to extend insurance coverage to all Americans? **LO22.6**

16. How does the PPACA attempt to ensure affordable health insurance for the poor? **LO22.6**

17. What were the objections made by opponents of the PPACA? **LO22.6**

18. **LAST WORD** What are the three major cost-reducing features of the Singapore health care system? Which one do you think has the largest effect on holding down the price of medical care in Singapore? What element of the Singapore system is shared by the Whole Foods and State of Indiana systems? What elements are missing? How difficult do you think it would be to implement those missing elements in the United States? Explain.

REVIEW QUESTIONS

1. Which of the following best describes the United States' level of health care spending as compared to that of other nations? **LO22.1**
- **a.** The lowest of all nations.
- **b.** A bit lower than average.
- **c.** Average.
- **d.** A bit higher than average.
- **e.** The highest of all nations.

2. Which of the following make a person *less* likely to have health insurance? **LO22.3**
*Select **one or more** answers from the choices shown.*
- **a.** Working for a larger firm.
- **b.** Being a low-wage worker.
- **c.** Being employed.
- **d.** Having excellent health.
- **e.** Being chronically ill.

3. A patient named Jen visits Dr. Jan. Dr. Jan is nearly certain that Jen only has a cold. But because Dr. Jan is afraid of malpractice lawsuits, she orders an extensive battery of tests just to make sure that Jen can never claim—if she turns out to have something more severe—that Dr. Jan shirked her duties as a medical professional. Dr. Jan's behavior is an example of: **LO22.4**
- **a.** Asymmetric information.
- **b.** Fee-for-service.
- **c.** Defensive medicine.
- **d.** Positive externalities.

4. All MegaCorp employees who stay on the job for more than three years are rewarded with a 10 percent pay increase and coverage under a private health insurance plan that MegaCorp pays for. Tina just passed three years as a MegaCorp employee and reacts to having health insurance by taking up several dangerous sports because now she knows that the insurance plan will pay for any injuries that she may sustain. This change in Tina's behavior is known as: **LO22.4**
- **a.** Defensive medicine.
- **b.** Asymmetric information.
- **c.** The moral hazard problem.
- **d.** The personal mandate.

5. By increasing demand, health insurance creates: **LO22.4**
- **a.** A deadweight loss related to overconsumption.
- **b.** A deadweight loss related to underconsumption.
- **c.** Neither of the above.

6. Ralph will consume any health care service just as long as its MB exceeds the money he must pay out of pocket. His insurance policy has a zero deductible and a 10 percent copay, so Ralph only has to pay 10 percent of the price charged for any medical procedure. Which of the following procedures will Ralph choose to consume? **LO22.5**
- **a.** An \$800 eye exam that has an MB of \$100 to Ralph.
- **b.** A \$90 hearing test that has an MB of \$5 to Ralph.
- **c.** A \$35,000 knee surgery that has an MB of \$3,000 to Ralph.
- **d.** A \$10,000 baldness treatment that has an MB of \$16,000 to Ralph.

7. True or False. Under the PPACA, Americans are free to decide for themselves whether or not they should have health insurance coverage. **LO22.6**

PROBLEMS

1. Suppose that the price elasticity for hip replacement surgeries is 0.2. Further suppose that hip replacement surgeries are originally not covered by health insurance and that at a price of \$50,000 each, 10,000 such surgeries are demanded each year. **LO22.2**
- **a.** Suppose that health insurance begins to cover hip replacement surgeries and that everyone interested in getting a hip replacement has health insurance. If insurance covers 50 percent of the cost of the surgery, by what percentage would you expect the quantity demanded of hip replacements to increase? What if insurance covered 90 percent of the price? (Hint: Do not bother to calculate the percentage changes using the midpoint formula given in Chapter 6. If insurance covers 50 percent of the bill, just assume that the price paid by consumers falls 50 percent.)
- **b.** Suppose that with insurance companies covering 90 percent of the price, the increase in demand leads to a jump in the price per hip surgery from \$50,000 to \$100,000.

How much will each insured patient now pay for a hip replacement surgery? Compared to the original situation, where hip replacements cost $50,000 each but people had no insurance to help subsidize the cost, will the quantity demanded increase or decrease? By how much?

2. The federal tax code allows businesses but not individuals to deduct the cost of health insurance premiums from their taxable income. Consider a company named HeadBook that could either spend $5,000 on an insurance policy for an employee named Vanessa or increase her annual salary by $5,000 instead. **LO22.4**

a. As far as the tax code is concerned, HeadBook will increase its expenses by $5,000 in either case. If it pays for the policy, it incurs a $5,000 health care expense. If it raises Vanessa's salary by $5,000, it incurs $5,000 of salary expense. If HeadBook is profitable and pays corporate profit taxes at a marginal 35 percent rate, by how much will HeadBook's tax liability be reduced in either case?

b. Suppose that Vanessa pays personal income tax at a marginal 20 percent rate. If HeadBook increases her salary by $5,000, how much of that increase will she have after paying taxes on that raise? If Vanessa can only devote what remains after paying taxes on the $5,000 to purchasing health insurance, how much will she be able to spend on health insurance for herself?

c. If HeadBook spends the $5,000 on a health insurance policy for Vanessa instead of giving it to her as a raise, how many more dollars will HeadBook be able to spend on Vanessa's health insurance than if she had to purchase it herself after being given a $5,000 raise and paying taxes on that raise?

d. Would Vanessa prefer to have the raise or to have HeadBook purchase insurance for her? Would HeadBook have any profit motive for denying Vanessa her preference?

e. Suppose the government changes the tax law so that individuals can now deduct the cost of health insurance from their personal incomes. If Vanessa gets the $5,000 raise and then spends all of it on health insurance, how much will her tax liability change? How much will she be able to spend on health insurance? Will she now have a preference for HeadBook to buy insurance on her behalf?

3. Preventive care is not always cost-effective. Suppose that it costs $100 per person to administer a screening exam for a particular disease. Also suppose that if the screening exam finds the disease, the early detection given by the exam will avert $1,000 of costly future treatment. **LO22.4**

a. Imagine giving the screening test to 100 people. How much will it cost to give those 100 tests? Imagine a case in which 15 percent of those receiving the screening exam test positive. How much in future costly treatments will be averted? How much is saved by setting up a screening system?

b. Imagine that everything is the same as in part *a* except that now only 5 percent of those receiving the screening exam test positive. In this case, how much in future costly treatments will be averted? How much is lost by setting up a screening system?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER 23

Immigration

Learning Objectives

LO23.1 Describe the extent of legal and illegal immigration into the United States.

LO23.2 Discuss why economists view economic immigration as a personal human capital investment.

LO23.3 Explain how immigration affects average wages, resource allocation, domestic output, and group income shares.

LO23.4 Relate how illegal immigration affects employment and wages in low-wage labor markets and impacts state and local budgets.

LO23.5 Demonstrate how economics can inform current immigration discussions and attempts to reform immigration laws.

The population of the United States is composed largely of immigrants and their descendants, yet immigration has long been a matter of heated controversy. Some immigration issues are political, social, and legal; others are economic. Our focus will be on economic issues and **economic immigrants**—international migrants motivated by economic gain. How many such immigrants come to the United States each year? What is their motivation and what economic impact do they make? Should more or fewer people be allowed to enter legally? What criteria, if any, should be used in allowing legal entry? How should the United States handle illegal immigration?[1]

[1]Some of our discussion is drawn from our more advanced treatment of mobility and migration in our textbook on labor economics: Campbell R. McConnell, Stanley L. Brue, and David A. Macpherson, *Contemporary Labor Economics,* 10th ed. (New York: McGraw-Hill, 2013), pp. 263–290.

Number of Immigrants

LO23.1 Describe the extent of legal and illegal immigration into the United States.

U.S. immigration consists of **legal immigrants**—immigrants who have permission to reside and work in the United States—and **illegal immigrants**—immigrants who arrive illegally or who enter legally on temporary visas but then fail to leave as stipulated. Legal immigrants include *permanent legal residents* ("green card" recipients) who have the right to stay in the country indefinitely and *temporary legal immigrants*, who have visas that allow them to stay until a specific date. Illegal immigrants are alternatively called unauthorized immigrants, illegal aliens, or, if working, undocumented workers.

Legal Immigrants

Figure 23.1 shows the annual levels of legal immigration into the United States since 1980. The spike in legal immigration from 1989 to 1991 resulted from an amnesty program through which many formerly illegal immigrants became legal residents. Between 2002 and 2011, legal immigration averaged 1 million a year. This number is higher than for earlier decades because beginning in 1990 the federal government increased the annual immigration quota from 500,000 to 700,000. (We provide a historical overview of U.S. immigration policy in this chapter's Last Word.)

Augmenting quota immigrants in some years are thousands of legal immigrants who are refugees (people who flee their country for safety) or are entrants to the United States under special provisions of the immigration law. As an example of the latter, the current **H1-B provision** of the immigration law allows 65,000 high-skilled workers in "specialty occupations" to enter and work continuously in the United States for six years. Such high-skilled occupations include high-tech workers, scientists, and professors.

A total of 1,062,040 people became permanent legal residents of the United States in 2011. About 55 percent of them were women and 45 percent were men. Around 57 percent of all legal immigrants were married.

As shown in Figure 23.2, about 65 percent of the 1,062,040 legal immigrants in 2011 were family-sponsored. They were parents, children, siblings, or other qualified relatives of legal permanent U.S. residents. Another 13 percent were admitted based on employment-based preferences. Most of these immigrants were sponsored by employers. Refugees, "diversity immigrants," and others accounted for the remaining 21 percent. The 50,000 quota for diversity immigrants is filled with qualified immigrants who are from countries with low rates of immigration to the United States. Because applications by diversity immigrants exceed the 50,000 quota, the slots are filled through an annual lottery.

Although the percentage varies somewhat each year, current U.S. immigration law is heavily weighted toward family reunification. This weighting is much heavier than that in Canada, which gives considerably stronger preference to immigrants with high levels of education and work skill. Of course, immigration by family ties and immigration by employment preferences are not necessarily mutually exclusive since a portion of the immigrants admitted through family ties are highly educated and skilled.

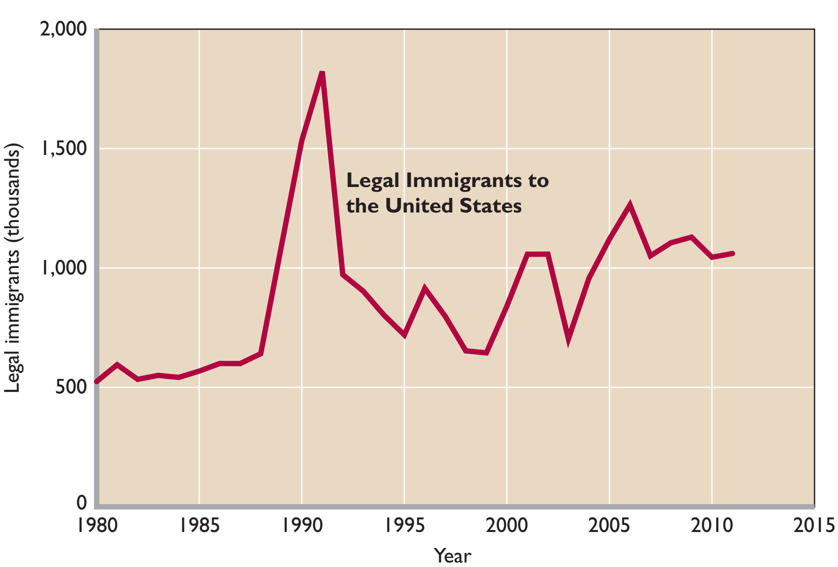
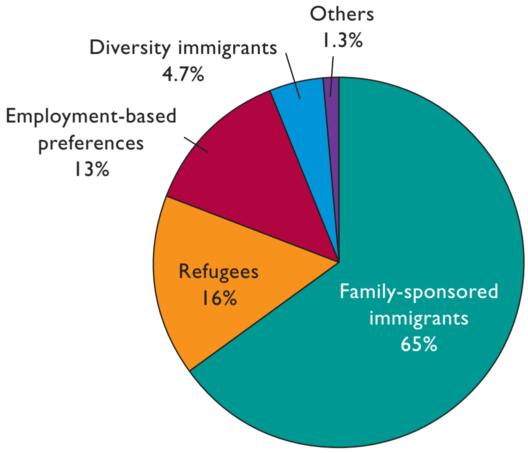


FIGURE 23.1 Legal immigration to the United States, 1980–2011. Legal immigration grew slowly between 1980 and 1988 and then spiked from 1989 to 1991 when previous illegal immigrants gained legal status as permanent residents under the terms of an amnesty program. Since that spike, legal immigration has remained relatively high, partly because the annual legal immigration quota was raised from 500,000 to 700,000. Within the totals are thousands of refugees, grantees of political asylum, and entrants under special provisions of the immigration law.

Source: Office of Immigration Statistics, Department of Homeland Security, **www.dhs.gov**.

FIGURE 23.2 Legal immigration by major category of admission, 2011. The large bulk of legal U.S. immigrants obtain their legal status via family ties to American residents.



Source: Office of Immigration Statistics, Department of Homeland Security, **www.dhs.gov**.

Table 23.1 shows the 10 leading countries of origin of U.S. legal permanent immigrants in 2011. Mexico topped the list with 143,446 immigrants, accounting for 14 percent of the total. China, the Philippines, and India were also heavy contributors to U.S. immigration in 2011. In recent years immigration has composed about one-third of the total growth of the U.S. population and one-half the growth of the U.S. labor force.

Illegal Immigrants

Our figures and tables do not include illegal immigrants. The U.S. Census Bureau estimates the number of illegal immigrants living in the United States through a residual approach. It finds the current total number of *all* immigrants through census surveys and then subtracts the sum of the past annual inflows of *legal* immigrants. The residuals are large. In fact, the net annual inflow of illegal immigrants averaged 250,000 per year from 2000 to 2009, with over 60 percent arriving from Mexico and Central America. As this chapter's Last Word describes, however, the net annual inflow reversed during and after the 2007–2009 recession.

Some illegal immigrants move back and forth across the U.S.-Mexican border, but in 2012 about 11.1 million illegal immigrants were residing continuously in the United States. Increasingly, illegal immigrants are continuous residents, not temporary workers who follow the agricultural harvest (as was typical several decades ago). An estimated 58 percent of the 11.1 million illegal immigrants originally came from Mexico, with many others arriving from nations spanning the globe. Illegal Mexican immigrants who work continuously in the United States work mainly outside of agriculture and, on average, have higher educational levels than Mexicans who do not migrate to the United States.

TABLE 23.1 U.S. Legal Immigrants by Top 10 Countries of Origin, 2011

Total	1,062,040
1. Mexico	143,446
2. China	87,016
3. India	69,013
4. Philippines	57,011
5. Dominican Republic	46,109
6. Cuba	36,452
7. Vietnam	34,157
8. South Korea	22,824
9. Colombia	22,635
10. Haiti	22,111

Source: Office of Immigration Statistics, Department of Homeland Security, *U.S. Legal Permanent Resident Report, 2011,* **www.dhs.gov**.

The Decision to Migrate

LO23.2 Discuss why economists view economic immigration as a personal human capital investment.

People immigrate into the United States (emigrate from their home countries) legally or illegally:

- To take advantage of superior economic opportunities.
- To escape political or religious oppression in their home countries.
- To reunite with family members or other loved ones, usually prior immigrants, who are already in the United States.

As previously stated, our interest is in economic immigration. Why do some workers uproot their lives to move from some other country to the United States? Why do other workers stay put?

Earnings Opportunities

The main driver of economic immigration is the opportunity to improve the immigrant's earnings and therefore standard of living. The chief attractor for economic immigrants is the availability of higher pay in the United States. In particular, immigrants can earn much higher wages in the United States than doing identical or nearly identical jobs in their home countries. Stated in terms of economic theory, immigrants reap larger financial rewards from their respective stocks of human capital when working in the United States rather than in their home countries.

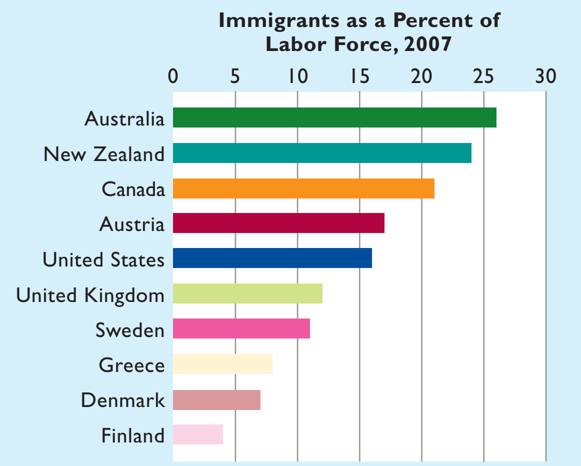
Recall that **human capital** is the stock of knowledge, know-how, and skills that enables a person to be productive, and thus to earn income. Other things equal, greater stocks of human capital (for example, more education or better training) result in greater personal productivity and earnings. But whatever someone's stock of education and skill, the value of that human capital depends critically on the capacity for it to earn income. That is where economic migration comes in. By securing higher earnings, migrants can increase—often quickly and dramatically—the value of their human capital. Economic migrants move from one country to another for the same reason many internal migrants move from one city or state to another within their home country: to increase their pay and therefore achieve a higher standard of living.

Other things equal, larger wage differences between nations strengthen the incentive to migrate and therefore increase the flow of immigrants toward the country providing the greater wage opportunities. Today, major "magnet countries" that attract a lot of immigrants include Australia, Switzerland, the United States, and several Western European nations. Global Perspective 23.1 shows the percentage of labor forces composed of foreign-born workers in selected nations in 2007, the latest year for which data are available. Along with earnings opportunities, significant differences in educational opportunities, health care availability, and public pensions and welfare benefits play a role in international migration decisions.

GLOBAL PERSPECTIVE 23.1

Immigrants as a Percent of the Labor Force, Selected Advanced Industrial Countries

Immigrants make up relatively large percentages of the labor forces in several advanced industrial countries, including Australia, Austria, and the United States, but not in other countries such as Finland.



Source: Organization for Economic Cooperation and Development, **www.oecd.org**.

Moving Costs

Immigration can be viewed as an investment decision. As with other investments, current sacrifices are necessary to achieve future benefits. In moving from one nation to another, workers incur personal costs. Some of these costs are explicit, out-of-pocket costs such as paying application fees (for example, $1,010 for a green card) and a number of moving expenses. For illegal immigrants, a major explicit cost may be a payment to an expediter—a "coyote"—who charges as much as $2,000 to smuggle someone into the United States and transport him or her to a major city such as Chicago, New York, Houston, or Los Angeles. Other costs of migrating are implicit. They are opportunity costs such as the income given up while the worker is moving and looking for a job in the new country. Still more subtle costs are incurred in leaving family and friends and adapting to a new culture, language, and climate. For illegal immigrants, there is the additional potential cost of being caught, jailed, and deported.

The prospective immigrant estimates and weighs all such costs against the expected benefits of the higher earnings in the new country. A person who estimates that the stream of future earnings exceeds the explicit and implicit costs of moving will migrate; a person who sees those costs as exceeding the future stream of earnings will stay put.[2]

Factors Affecting Costs and Benefits

Earnings differences provide the major incentive to migrate, but many nonwage factors also affect the cost-benefit evaluation of moving to another nation. Let's look at two key factors.

Distance Other things equal, greater distance reduces the likelihood of migration. Most obvious, transportation costs rise with distance. In addition, migration to more distant countries is often seen as riskier because information

[2]As with other investment decisions, the decision to move internationally requires a comparison of the *present value* of the stream of additional earnings and the *present value* of the costs of moving. Present value considerations (discussed in Chapters 16 and 17) complicate the decision to migrate but do not alter the basic analytical framework.

about job market conditions in distant countries is usually less certain than about job market conditions in nearby countries (although the Internet has greatly reduced this difference in recent years). Finally, the farther the move, the greater the possible costs of maintaining contact with friends and family. In terms of expense, a short trip back across a border by automobile is one thing; an airline flight to a different continent, quite another.

The majority of international migrants move to countries relatively close to their home countries. Most Mexican migrants move to the United States. The majority of Eastern European migrants move to Western Europe. Close proximity reduces the cost of the move relative to anticipated benefits.

Some migrants, of course, *do* move to far-away lands. They often reduce their costs of these long moves by following **beaten paths**—routes taken previously by family, relatives, and friends. They also tend to cluster, at least for awhile, in cities and neighborhoods populated by former and current immigrants. For example, thousands of Russian immigrants have located in Brooklyn, New York. Numerous Asian immigrants have located in San Francisco.

The earlier immigrants ease the transition for those who follow by providing job information, employment contacts, temporary living quarters, language help, and cultural continuity. That reduces the costs and increases the benefits of migration to those who follow. Eventually, some new immigrants exploit even better economic opportunities by moving from the original place of migration to other cities within the country to which they have migrated. New clusters of immigrants emerge and new migration networks result. For instance, these new clusters and networks help explain the rapid northward expansion of Latino immigrants during the past decade.

Age Younger workers are much more likely to migrate than older workers. Age affects both benefits and costs in the calculation to move or stay put. Particularly relevant, younger migrants have more years to recoup their costs of moving. Spread over decades, the higher wage in the new country builds to a large accumulation of additional earnings relative to the earnings that would have accrued if the person had not moved. In contrast, older people are closer to retirement and therefore may conclude that moving abroad simply is not worth the effort. Their added earnings over their remaining work years simply will not be sufficient to cover the costs of the disruption and move.

Younger migrants also tend to have lower moving costs than older workers. For example, they often have accumulated fewer personal possessions to transport. Younger workers generally have fewer roots and ties to the local community and so may find it easier to adapt to new customs and cultures. This greater flexibility reduces the perceived costs of moving and increases the likelihood that the younger person will move.

Younger workers are also more likely to be single or, if married, less likely to have started families than older workers. The potential costs of migrating multiply rapidly when spouses and families are present. Finding affordable housing large enough for families and enrolling children in new schools complicate the potential move. Also, when both spouses work, the secondary wage earner may face a period of unemployment before finding a new job in the new country.

Other Factors Several other factors may affect the cost and benefit calculations of immigrants into the United States. Studies show that immigrants who lack English language skills do not, in general, fare as well in the U.S. workforce as immigrants who have those skills when they arrive. For some highly skilled immigrants, lower tax rates or opportunities to set up businesses in the United States may be the draw. Also, some immigrants to the United States may be willing to endure low or even negative personal returns on immigration simply so that their children have greater economic opportunities than at home.

QUICK REVIEW 23.1

- An average of 1 million legal immigrants and 250,000 illegal immigrants entered the United States each year between 2002 and 2011.
- In 2011 Mexico was the greatest single contributor (14 percent) to U.S. legal immigration. Hundreds of thousands of additional legal immigrants arrived from China, the Philippines, India, and many other nations.
- Economists view economic migration as a personal investment; a worker will move internationally when the expected gain in earnings exceeds the explicit and implicit costs of moving.
- Other things equal, the greater the migration distance and the older the prospective migrant, the less likely the person will move.

Economic Effects of Immigration

LO23.3 Explain how immigration affects average wages, resource allocation, domestic output, and group income shares.

Immigration creates personal gains for movers, and also affects wage rates, efficiency, output, and the division of income. Like international trade, migration produces large economic benefits but also creates short-term winners and losers. In particular, we will see that the wage-rate and division-of-income aspects of immigration are two main sources of controversy.

Personal Gains

The fact that economic immigration to the United States is sizeable and continuous affirms that in general the economic benefits of immigration to the immigrants exceed their costs. In economic terms, the inflows of legal and illegal immigration indicate that this investment has a positive return to movers. Studies confirm that the returns to immigrating to the United States are, on average, quite substantial. This should not be a surprise. For example, the real wages earned by recent Mexican male migrants to the United States are as much as six times higher than those earned by similarly educated men in Mexico.

Nevertheless, not all economic immigrants to the United States succeed. Migration decisions are based on expected benefits and are made under circumstances of uncertainty and imperfect information. High average rates of return do not guarantee that all migrants will benefit. In some cases the expected gain from immigration does not materialize—the anticipated job is not found in the new country, the living costs are higher than anticipated, the anticipated raises and promotions are not forthcoming, the costs of being away from family and friends are greater than expected. Major **backflows**—return migration to the home country—therefore occur in most international migration patterns, including those between the United States and other countries.

Although this return migration may be costly to those involved, it increases the availability of information about the United States to other potential migrants. These people are then able to better assess the benefits and costs of their own potential moves.

Also, although economic immigrants on average improve their standard of living, they may or may not achieve pay parity with similarly educated native-born workers. The skills that migrants possess are not always perfectly transferable between employers in different countries because of occupational licensing requirements, specific training, or differing languages. This lack of **skill transferability** may mean that migrants, although improving their own wage, may earn less than similarly employed native-born workers in the United States. Studies find that this is particularly true for immigrants who lack English-language skills.

On the other hand, a great deal of economic migration is characterized by **self-selection.** Because migrants choose to move while others with similar skills do not, it is possible that those who move possess greater motivation for personal economic gain and greater willingness to sacrifice current consumption for higher levels of later consumption. If so, these migrants may overcome the problem of imperfect skill transferability and eventually outdo domestic-born workers in wage and salary advancement. This possibility is particularly true of highly skilled immigrants such as scientists, engineers, physicians, and entrepreneurs.

Many U.S. firms have been founded by immigrants. In fact, immigrant entrepreneurs have founded about 35 percent of U.S. semiconductor companies, 31 percent of U.S. computer/communications companies, and 27 percent of U.S. software companies.

CONSIDER THIS . . .

Stars and Stripes

Recent immigrants have contributed mightily to the vitality of the American economy. The Council of Economic Advisers reminds us that:

> Skilled migrants, whether permanent or temporary, enrich our scientific and academic communities, boost the technical capabilities of U.S. firms (and the native-born workers employed there), augment the supply of health-care providers, and pay more in taxes than they absorb in government services. Many of these workers were educated in American universities and nearly all adjust easily to life in the United States in terms of language skills and employment. They make major innovative contributions in science, medicine, and engineering, and help keep the United States at the forefront of technological capability.*

Since 1990, U.S. immigrants have founded 1 of every 4 public companies backed by venture capital. Yahoo!, Intel, eBay, Google, and Sun Microsystems all have one or more immigrant founders. About half the engineers and people with computer science doctorates in the United States are foreign-born. One-fourth of U.S. major league baseball players were born abroad. Children of immigrants often dominate U.S. math and science competitions. And our list could go on.†

*_Economic Report of the President, 2007,_ p. 201.

†Data are from the National Foundation for American Policy, **www.nfap.net**.

Impacts on Wage Rates, Efficiency, and Output

Although the personal outcomes of immigration are relatively straightforward and easy to understand, the broader economic outcomes are somewhat complicated and obscure. A simple economic model of migration will help us sort through key cause-effect relationships and identify

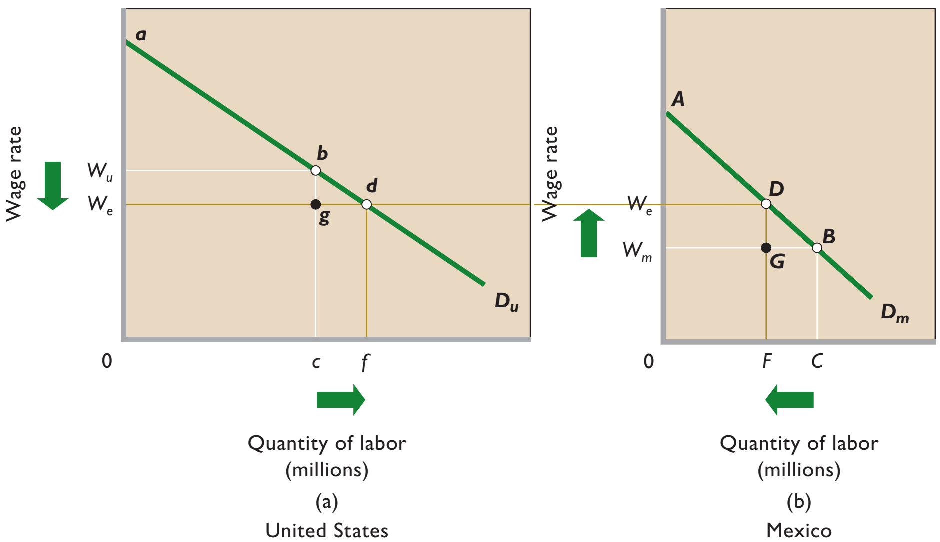


FIGURE 23.3 A simple immigration model. (a) The migration of low-wage Mexican labor by the amount *cf* to the United States increases U.S. domestic output by *cbdf*, reduces the U.S. wage rate from W_u to W_e, and increases U.S. business income by W_eW_ubd. (b) The out-migration of labor of *CF* from Mexico reduces Mexican domestic output by *FDBC*, raises the wage rate from W_m to W_e, and lowers Mexican business income by W_mW_eDB. Because the U.S. gain in domestic output of *cbdf* exceeds Mexico's loss of domestic output of *FDBC*, the migration depicted increases economic efficiency and produces a net gain of world output.

broader economic outcomes. In Figure 23.3a, D_u is the demand for labor in the United States; in Figure 23.3b, D_m is the demand for labor in Mexico. The demand for labor presumably is greater in the United States because it has more capital, advanced technology, and better infrastructure that enhance the productivity of labor. (Recall from Chapter 14 that the strength of labor demand is based on the marginal revenue productivity of labor.) Conversely, labor demand in Mexico is weaker since machinery and equipment are less abundant relative to labor, technology is less advanced, and infrastructure is less developed. We also assume that the before-migration labor forces of the United States and Mexico are *c* and *C*, respectively; that neither country is experiencing substantial long-term unemployment; and that labor quality in the two countries is the same.

If we further suppose that migration (1) has no cost, (2) occurs solely in response to wage differentials, and (3) is unimpeded by law in both countries, then workers will migrate from Mexico to the United States until wage rates in the two countries are equal at W_e. At that level, $C - F$ (equals $f - c$) workers will have migrated from Mexico to the United States. Although the U.S. wage level will fall from W_u to W_e, domestic output (the sum of the marginal revenue products of the entire workforce) will increase from 0*abc* to 0*adf*. This domestic output is the total output produced within the borders of the United States and equals U.S. domestic income.

In Mexico, the wage rate will rise from W_m to W_e, but domestic output will decline from 0*ABC* to 0*ADF*. Observe that the gain in domestic output *cbdf* in the United States exceeds the loss of domestic output *FDBC* in Mexico. The migration from Mexico to the United States has clearly increased the world's output and income.

The elimination of barriers to the international flow of labor tends to create worldwide **efficiency gains from migration.** The same number of workers—rearranged among countries—produces greater total output and income after migration than before migration. The world gains output (and income) because the freedom to migrate enables people to move to countries where they can contribute more to world production. Economic migration not only provides a positive investment return to the mover, it produces an overall efficiency gain. Migration enables the world to produce a larger output with its currently available resources. So labor mobility joins capital mobility and international trade in enhancing the world's standard of living.

Income Shares

With personal gains and overall productivity gains, why would anyone oppose immigration? Our graphical model helps answer that question. There are specific groups of gainers and losers from immigration in both nations.

In Figure 23.3, as workers move from Mexico to the United States in search of higher wages, U.S. output will increase while Mexican output will decrease. These gains partly explain why the United States encourages a relatively high level of immigration through high annual quotas. It also explains why some countries try to discourage outflows of labor from their countries. In particular, countries are rightfully concerned about the emigration of highly educated workers, particularly when those citizens received subsidized education at home. Such undesirable outflows are commonly called **brain drains.**

Figure 23.3 reveals a second consequence of immigration on income shares. The decline in the wage rate from

W_u to W_e in the United States reduces the wage income of native-born U.S. workers from $0W_ubc$ to $0W_egc$. The opposite outcome occurs in Mexico, where the average wage for the native-born workers who do not migrate rises.

Although we can specify income gains and losses to domestic-born workers, we cannot say what will happen to the total wage income (= domestic-born wage income + immigrant wage income) in each country. That depends on the elasticities of labor demand. For example, if labor demand is elastic, the wage decrease in the United States will increase total wage income. In contrast, if labor demand is inelastic, the same wage decrease will cause total wage income to fall.

The immigration-caused decline in wage income for native-born U.S. workers is a major reason that many U.S. labor unions oppose increasing immigration quotas in the United States. Unions tend to resist policies that reduce the wages of their current membership or undercut their bargaining power by creating larger pools of potential workers for nonunion firms. In direct contrast, the increase in wages in the outflow country is a possible reason why labor groups in Mexico show little concern about the large-scale outflow of Mexican labor to the United States.

Finally, Figure 23.3 shows that the immigration enhances business income in the United States while reducing it in Mexico. The before-immigration domestic output and income in the United States is represented by area $0abc$. Total wage income is $0W_ubc$—the wage rate multiplied by the number of workers. The remaining triangular area W_uab shows business income before immigration. The same reasoning applies to Mexico, where the triangle W_mAB represents before-immigration business income.

Unimpeded immigration increases business income from W_uab to W_ead in the United States and reduces it from W_mAB to W_eAD in Mexico. Other things equal, owners of U.S. businesses benefit from immigration; owners of Mexican businesses are hurt by emigration. These outcomes are what we would expect intuitively; the United States is gaining "cheap" labor and Mexico is losing "cheap" labor. This conclusion is consistent with the historical fact that U.S. employers have often actively recruited immigrants and have generally supported higher immigration quotas, liberal guest-worker programs, and expanded specialized work visas such as H1-Bs.

Complications and Modifications

Our model is a purposeful simplification of the much more complex actual reality. Thus, it is not surprising that the model includes simplifying assumptions and overlooks some factors that may be important in certain situations. Relaxing some of the assumptions and introducing the omitted factors may affect our conclusions.

Costs of Migration Our model assumed that the movement of workers from Mexico to the United States is without personal cost, but we know that migrants incur explicit, out-of-pocket costs of physically moving and the implicit opportunity costs of forgone income during the move and transition.

In Figure 23.3, the presence of migration costs means that the flow of labor from Mexico to the United States will stop short of that required to close the wage differential entirely. Wage rates will remain somewhat higher in the United States than in Mexico, and that wage-rate difference will not encourage further migration to close up the wage gap. At some point, the remaining earnings gap between the two countries will not be sufficient to cover the marginal cost of migration. Migration will end, and the total output and income gain from migration will be less because wages have not equalized.

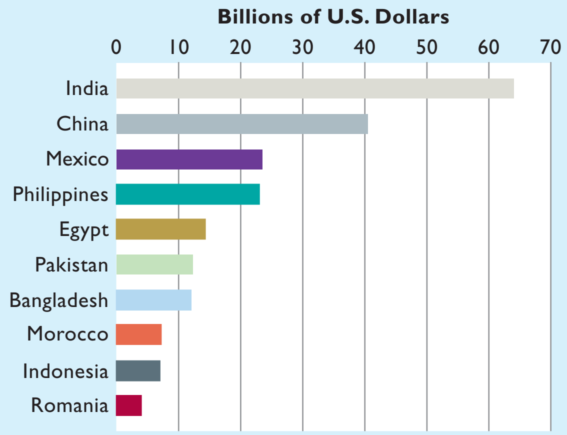
Remittances and Backflows Although most of the workers who acquire skills in the country of immigration do not return home, some migrants see their moves as temporary. They move to a more highly developed country; accumulate some wealth, training, or education through hard work and frugality; and return home to establish their own enterprises. During their time in the new country, these and other migrants frequently make sizable **remittances** to their families at home. These money transfers to the home country redistribute the net gain from migration between the countries involved.

In Figure 23.3, remittances by Mexican workers in the United States to their relatives in Mexico would cause the gain in U.S. income retained in the United States to be less than the domestic output and income gain shown. Similarly, the loss of income available in Mexico would be less than the domestic output and income loss shown. The World Bank estimates that $24 billion of remittances—an amount equal to 2 percent of Mexico's GDP—flowed to Mexico from other countries in 2011. Most of these remittances originate in the United States and are a major reason Mexico favors liberal U.S. immigration laws and generally opposes U.S. policies to stem the flow of illegal immigrants across the U.S. border. (Global Perspective 23.2 shows selected developing countries receiving remittances by emigrants in 2011. For many developing countries, remittances exceed foreign direct investment—purchases by foreigners of lasting ownership interests in domestic firms—as a source of foreign currencies available to buy imported goods and services.)

GLOBAL PERSPECTIVE 23.2

Emigrant Remittances, Selected Developing Countries, 2011

Although both developing nations and advanced industrial nations receive remittances from their emigrants, most remittances flow toward developing nations. For some of these nations, the amount of remittances each year exceeds their direct foreign investment (economic investment by foreign individuals and firms).



Source: World Bank, **www.worldbank.org**.

Along with remittances, backflows of migrants to their home countries might also alter gains and losses through time. For example, if some Mexican workers who migrated to the United States acquired substantial labor market or managerial skills and then returned home, their enhanced human capital might contribute substantially to economic development in Mexico. Further, some of the more successful U.S. immigrants eventually may use their expertise and wealth to help build new businesses in Mexico. Both will eventually increase labor demand in Mexico and raise wage rates there.

Complementary versus Substitute Resources Although the average wage rate of domestic-born workers may decline because of immigration, not all such workers will see their wages fall. Many immigrant workers and domestic-born workers are **complementary resources** rather than **substitute resources** (Chapter 14). When that is the case, the lower wage rate resulting from large-scale immigration reduces production costs, creating an output effect that raises labor demand for certain domestic-born workers. For example, the large number of immigrants working in the home building industry lowers construction wages and reduces the cost of home building. That in turn increases the number of houses built and sold, which increases the demand for domestic-born residents who help manufacture sheet rock, plumbing products, air conditioners, major appliances, and other home products.

Expansion of Capital Long-run effects on capital are another reason native-born citizens may not be permanently harmed to the extent suggested by the simple immigration model. The stock of capital was implicitly constant in both countries in Figure 23.3, fixing the demand curves in place. But the rise in business income in the United States relative to the stock of capital produces a higher rate of return on capital. The higher rate of return stimulates overall investment, which in the long run adds to the size of the nation's stock of capital. Normally, the addition of new capital such as plant and equipment raises labor productivity, lowers production costs, and reduces product prices. As a result, wages and salaries rise because of increased demand for labor.

On the other hand, the inflow of illegal workers into certain low-wage occupations such as field harvesting may stifle R&D, technological advance, and investment in some industries. The easy availability of inexpensive legal or illegal immigrant labor provides little incentive to mechanize or otherwise economize on the use of labor. In this regard, economists note that the temporary slowing of the flow of illegal agricultural workers after the terrorist attacks of September 11, 2001, increased the purchase of mechanical harvesting equipment such as tree-trunk shakers used to harvest oranges.

Full Employment versus Unemployment Our model conveniently assumes full employment in both countries. Mexican workers presumably leave low-paying jobs to take higher-paying jobs in the United States (more or less immediately). However, in many circumstances, the factor that pushes immigrants from their homelands is not simply low wages but chronic unemployment or underemployment. Many developing countries have large populations and surplus labor. A sizeable number of workers are either unemployed or so grossly underemployed that their contribution to domestic output is zero or near zero.

If we allow for this possibility, then Mexico would gain rather than lose by having such workers emigrate. Unemployed Mexicans are making no or little contribution to Mexico's domestic output and must be sustained by transfers from the rest of the labor force. The remaining Mexican labor force will be better off by the amount of the transfers after the unemployed workers have migrated to the United States.

The unemployed workers moving to the United States may reflect **negative self-selection,** in which movers are less capable and perhaps less motivated than similarly educated people who did not immigrate. This possibility could conceivably join higher domestic wages and large remittances as an explanation of why Mexico generally opposes stronger border enforcement by the United States.

Conversely, if the Mexican immigrant workers are unable to find jobs in the United States and are sustained through transfers from employed U.S. workers, then the after-tax income of working Americans will decline. This fear is one reason many Americans oppose immigration of low-education, low-skilled workers to the United States.

Fiscal Impacts

What effects do immigrants have on tax revenues and government spending in the United States? Do they contribute to U.S. GDP, as our model suggests, or do they go on welfare and use "free" public goods, draining the government treasury?

Before the 1970s, the immigrant population was less likely to receive public assistance than people born in the United States. Migrants were typically young, single men with significant education and job training. They were readily employable in average-paying jobs and therefore were net contributors to the tax-expenditure system.

But the situation reversed between the 1970s and 1998, when immigrants began to use the welfare system proportionately more than natives. The changing mix of immigrants from relatively skilled workers toward unskilled workers explained the turnabout. Critics claimed that U.S. welfare programs were drawing unskilled (and often illegal) workers to the United States from some of the world's poorest nations. Immigrants made up more than 10 percent of Supplemental Security Income (SSI) rolls in 1998 compared to only 3.3 percent a decade before.

As a result of this trend, the major overhaul of the U.S. welfare system that was passed into law in 1996 denied welfare benefits to new legal immigrants for their first five years in the United States. As a result, between 1996 and 2006 cash welfare payments to immigrants declined by 73 percent, food stamps by 39 percent, and SSI payments by 20 percent.

It remained the case, however, that rates of welfare utilization continued to be higher among immigrants than nonimmigrants. In 2009, for instance, a survey of all households with minor children revealed that 57 percent of immigrant households used at least one form of welfare as compared with only 39 percent of nonimmigrant households. Among households headed by an illegal immigrant, the percentage stood even higher, at 71 percent.

In addition to utilizing welfare at high rates, low-income immigrants impose costs on state and local governments by enrolling children in public schools, using emergency health care facilities, and straining the criminal justice system. For low-income immigrants, these fiscal burdens substantially exceed taxes paid. We will say more about this later.

Research Findings

All economists agree that U.S. immigration increases U.S. domestic output and income and that highly educated immigrants and successful entrepreneurs add to the vitality of American enterprise. But in light of the complications just discussed, no single generalization is possible as to the impact of immigration on the wages of native-born U.S. workers.

The best evidence indicates that immigration reduces the wages of native-born workers who have low educations, and also may reduce the salaries of some highly trained native-born workers. For example, studies show that immigration reduces the wages of native-born Americans who do not have high school diplomas, native-born African-American men, and native-born holders of doctorate degrees.

The overall effect of immigration on the average American wage is much less clear. Scholarly estimates on that effect range from minus 3 percent to plus 2 percent.[3]

QUICK REVIEW 23.2

- Other things equal, immigration reduces the average wage rate, increases domestic output, lowers the total wage income of native-born workers, and bolsters business income in the destination nation; it has the opposite effects in the origin nation.
- Assessing the impacts of immigration is complicated by such factors as remittances and backflows, complementary versus substitute labor, investment impacts, unemployment, and fiscal effects.

[3]The research conclusions summarized in this section are based on recent studies conducted individually or jointly by several prominent economists, including George Borjas, David Card, Richard Freeman, Jeffrey Grogger, Gordon Hanson, Lawrence Katz, Gianmarco Ottaviano, and Giovanni Peri.

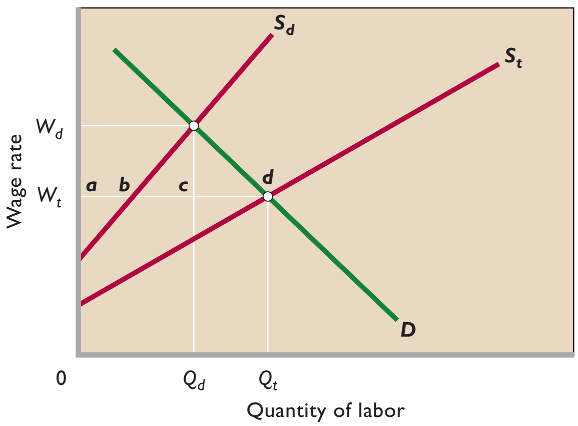
The Illegal Immigration Debate

LO23.4 Relate how illegal immigration affects employment and wages in low-wage labor markets and impacts state and local budgets.

Much of the recent concern about immigration has focused on illegal immigration, not immigration per se. Economists point out that a strong inflow of undocumented workers to some extent reflects the increasing scarcity of domestic unskilled labor in the United States. Only about 12 percent of the native-born U.S. workforce has less than a high-school diploma today, compared to about 50 percent in 1960. That scarcity has created significant employment opportunities for unskilled illegal immigrants. Illegal workers make up roughly 25 percent of all agricultural workers, 19 percent of all cleaning workers, 17 percent of construction workers, and 12 percent of food preparers.

Many Americans fear that illegal immigrants and their families depress wage rates in these and other already low-wage U.S. occupations and also burden American citizens through their use of public services such as emergency medical care and public schools. Are these concerns justified?

FIGURE 23.4 The impacts of illegal workers in a low-wage labor market. Illegal workers in a low-wage labor market shift the labor supply curve, as from S_d to S_t, and reduce the market wage from W_d to W_t. At wage W_t, *ab* workers are domestic-born (or legal residents) and *bd* workers are illegal immigrants. If all the illegal workers were deported, however, Q_d American workers would be employed. To say that illegal workers do jobs that Americans are not willing to do (at any wage rate), therefore, is somewhat misleading. Similarly misleading is the conclusion that the deportation of illegal workers would boost the employment of American workers on a one-for-one basis.



Employment Effects

Two extreme views on illegal immigration are often expressed. Some observers suggest that the employment of illegal workers decreases the employment of legal workers on a one-for-one basis. They erroneously suggest that the economy has only a fixed number of jobs at any time. Supposedly, every job taken by an illegal worker deprives a legal resident of that job. At the other extreme is the claim that illegal workers accept only work that legal residents will not perform. Viewed this way, illegal workers displace *no* legal residents from their jobs.

Both views are misleading. Consider Figure 23.4, which illustrates a market for unskilled field workers in agriculture. The downsloping curve *D* is the labor demand curve for field workers. The upsloping supply curve S_d is the labor supply of domestic-born workers, while curve S_t reflects the combined total supply of domestic-born workers and illegal immigrants. The horizontal distances between S_t and S_d at the various wage rates measure the number of illegal immigrants offering their labor services at those wage rates.

With illegal workers present, as implied by curve S_t, the equilibrium wage and level of employment in this labor market are W_t and Q_t. At the low wage of W_t, only *ab* domestic-born workers are willing to work as field hands; the other workers—*bd*—are illegal immigrants. The low employment of domestic-born workers presumably is caused by their better wage opportunities and working conditions in alternative occupations or by the availability of government transfer payments. Recall that illegal workers are not eligible for most welfare benefits.

Can we therefore conclude from Figure 23.4 that illegal workers have filled field jobs that most U.S.-born workers do not want? The answer is "yes," but only with the proviso: "at wage rate W_t." With fewer illegal immigrants in this labor market, labor supply would be less than that shown by curve S_t. The wage rate would be higher than W_t, and more legal residents would offer their services as field hands. For example, if the United States cut off the full inflow of illegal workers to this market, the relevant supply curve would be S_d and the wage rate would rise to W_d. Then Q_d domestic-born workers as opposed to *ab* workers would work as field hands. The critical point is that the willingness of Americans to work at any particular job depends significantly on the wage rate being paid. A sufficiently high **compensating wage differential** (wage premium to compensate for undesirable work) will attract U.S. workers even to otherwise undesirable work.

The opposite argument, that illegal workers reduce the employment of Americans by an amount equal to the employment of illegal workers, is also misleading. Figure 23.4 reveals that the illegal workers increase the total number of jobs in the labor market. With illegal workers, the number of jobs is Q_t. Without those workers, it is only Q_d. The deportation of the illegal workers would not increase domestic employment on a one-for-one basis. Native-born employment would increase by the amount *bc* in this specific labor market, not by *bd*.

Generally, illegal immigration causes some substitution of illegal workers for domestic workers, but the amount of displacement is less than the total employment of the illegal workers. Illegal immigration—as with legal immigration—increases total employment in the United States.

Wage Effects

Large flows of illegal workers into specific low-wage labor markets reduce wage rates in those markets. Note in Figure 23.4 that the greater supply of field workers reduces their wage rate from W_d to W_t. Some U.S. wages—including those of field laborers, food preparers, and house cleaners—are lower than otherwise because of illegal immigration.

As discussed previously, the overall effect of illegal immigration on the average wage rate in the economy is either a smaller decline or even positive. As with legal immigrants, some illegal workers are complementary inputs to domestic-born workers, not substitutes. An example of this complementarity would be illegal fruit pickers and the domestic-born truck drivers who deliver the fruit to grocery stores. The lower price of the fruit increases the amount of fruit demanded and thus the amount of it that needs to be delivered. That increases the labor demand for the complementary truck drivers, whose wage rates rise.

Only where illegal workers and legal workers are substitute resources will the increase in labor supply reduce the wages of other workers. Ironically, studies show that the largest negative impact of illegal immigrants is on the wages of previous immigrants, not on those of native-born workers.

Illegal immigration has very little effect on the average level of wages in the United States. That average wage level depends mainly on the nation's overall level of labor productivity, which illegal immigration does not appreciably affect.

Price Effects

Because illegal immigrants work at lower pay than would be necessary to attract native-born workers, the prices of goods and services that illegal workers produce are lower than they would be otherwise. The extent of such price reduction depends on several factors, including how much of the total cost of producing and delivering a product involves the services performed by illegal immigrants. In industries where illegal immigrants are heavily used—for example, construction, agriculture, landscaping, home cleaning, restaurant meals, and lodging—the presence of illegal workers may have a discernable downward price effect. Lower prices raise the standard of living of all Americans and their families.

Fiscal Impacts on Local and State Governments

One major and very legitimate concern about illegal immigration is the negative fiscal impact it has on local and state governments. Cities and states with high concentrations of illegal immigrants bear the main burden. The federal government receives the payroll taxes and income taxes withheld from the earnings of some illegal immigrants, but the state and local governments bear most of the costs of their presence. Immigrants place their children in local schools, use local emergency medical care, and add to the cost of the criminal justice system, most of which is provided by state and local governments. Immigrants do, however, pay state sales taxes and taxes on gasoline, and indirectly pay property taxes built into rent.

The average fiscal burden (government benefits minus taxes paid) on state and local government for each low-skilled immigrant household may be as high as $19,500 per household per year. In 2006 about 40 percent of the 4.5 million households falling into this low-skilled category were headed by illegal immigrants. One recent estimate of the fiscal burden for these households as a group is nearly $50 billion annually.

Other Concerns

Critics of illegal immigration point to other reasons to be concerned about illegal immigration. First, they say that allowing immigrants to enter the United States unlawfully undermines general respect for the law. If immigration laws can be broken, why can't other laws also be broken? The success of many immigrants in entering the United States and working for employers illegally rests on other criminal activity such as the creation of fake birth certificates, Social Security cards, and drivers' licenses. Also, some unauthorized immigrants engage in illegal activities such as drug smuggling, identity theft, and insurance fraud. Although U.S. immigrants (legal and illegal) have considerably lower prison rates than the native-born U.S. population, the crime rate for illegal immigrants is much higher than that of the native-born U.S. population.

Second, critics of the ineffective enforcement of border and employment laws point out that illegal immigration is highly unfair to the thousands of people enduring the expense and long waits associated with the process for legally gaining the right to live and work in the United States.

Finally, some observers see national defense as the greatest long-term risk from porous borders. The flow of illegal entrants into the United States is clearly at odds with the goal of homeland security. Ineffective border enforcement against illegal immigrants allows career

LAST WORD

The Startling Slowdown in Illegal Immigration

Illegal Immigration Fell Precipitously during the 2007–2009 Recession. Why Did It Happen? And Will It Continue?

The number of illegal immigrants living in the United States more than tripled between 1990 and 2007, increasing from 3.5 million in 1990 to 12.0 million in 2007. But during the 2007–2009 recession, the trend reversed and the number of illegal immigrants living in the United States fell by nearly 8 percent, down to 11.1 million in 2009.

The decline occurred because the backflow of illegal immigrants returning to their home countries exceeded the inflow of illegal immigrants entering the United States. The most important underlying cause for the backflow exceeding the inflow was quite simple: decreased job prospects for illegal immigrants.

There had been a major boom in housing prices and new home construction during the early 2000s, but it peaked in 2006 before crashing in 2007. When it did, the demand for unskilled labor fell dramatically. The ensuing recession cut the demand for unskilled labor even further.

As a result, most economists concluded that illegal immigration had declined due to low employment demand. Following the same logic, they also assumed that when the economy picked up again, so would illegal immigration.

Thus, it was a major surprise when the overall size of the illegal immigrant population stayed steady at 11.1 million in both 2010 and 2011. Economic growth had returned, but the net flow of illegal immigrants remained at about zero.

Economists have come up with several explanations for this leveling-off in the overall number of illegal immigrants. Taken together, they offer a plausible explanation for why illegal immigration remained subdued in 2010 and 2011—and for why it may decrease in coming years.

The first major factor has been the rapid decline in birthrates in the countries that have historically sent the most illegal immigrants to the United States. Consider Mexico. In 1960, Mexican women were having, on average, 7.3 births per woman per lifetime. By 2009, that number had tumbled to an average of just 2.4 births per lifetime—and, indeed, that number is expected to fall even further in coming years. As a result, the number of young Mexicans entering the job market each year in Mexico is now far lower than it was just a decade ago. One result has been far fewer young Mexicans deciding to seek their fortunes in the United States.

At the same time, economic growth has picked up substantially in most of the countries that have traditionally sent the most illegal immigrants to the United States. Consider Mexico once again. Mexico's economy grew nearly 5 percent per year between 2009 and 2011. That rapid growth drove up local wages—thereby making the wages available in the United States seem relatively less attractive than they had seemed just a few years earlier.

Government policies have also played a role. The North American Free Trade Agreement (NAFTA) that was signed by the governments of Canada, Mexico, and the United States in 1992 created a continent-wide free trade zone that spurred the creation of a larger manufacturing sector in Mexico. Mexico passed China and South Korea to become the world's largest producer of flat-screen TVs in 2009 and, in addition, Mexico has also become one of the world's top car and truck producers. Thus, many Mexicans now work in Mexican factories producing goods for export. Because they are employed in good-paying jobs in Mexico, they are less likely to want to immigrate illegally to the United States.

U.S. immigration-enforcement activities have also increased in recent years. A fence was built along several hundred miles of the U.S.-Mexican border and border patrols were strengthened after the 9/11 terrorist attack. As a result, the number of illegal immigrants deported by the U.S. Department of Homeland Security each year more than doubled between 2001 and 2011, rising from 189,026 removals in 2001 to 391,953 removals in 2011.

The factors just discussed—falling birthrates, better local job opportunities, and stronger border enforcement—lead many economists to conclude that illegal immigration is more likely to decrease than increase in coming years. If so, the debate over illegal immigration will likely take a less-prominent place in American political life. Whether one believes illegal immigration to be a serious problem or a net benefit, there will be less to argue over if the number of illegal immigrants declines in coming years.

criminals and even terrorists to enter the United States undetected.

QUICK REVIEW 23.3

- Illegal immigrants reduce wage rates in low-wage labor markets, take jobs that some Americans do not want, and expand total employment in low-wage occupations.
- The deportation of illegal immigrants would increase the wage rate in low-skilled labor markets but not increase employment on a one-to-one basis with the number of illegal workers deported.
- Illegal immigration imposes a high net fiscal burden on state and local governments.

Optimal Immigration

LO23.5 Demonstrate how economics can inform current immigration discussions and attempts to reform immigration laws.

The immigration issues relating to quotas and illegal immigrants go well beyond economics. They are also political and cultural issues. Nevertheless, economics can help inform the debate. Economic analysis suggests that immigration can either benefit or harm a nation, depending on the number of immigrants; their education, skills, and work ethic; and the rate at which they can be absorbed into the economy without disruption.

From a strictly economic perspective, immigration should be expanded until its marginal benefit equals its marginal cost. The MB = MC conceptual framework explicitly recognizes that there can be too few immigrants, just as there can be too many. Moreover, it recognizes that from a strictly economic standpoint, not all immigrants are alike. Some immigrants generate more benefits to the U.S. economy than others; and some immigrants impose more costs on taxpayers than others. The immigration of, say, a highly educated scientist obviously has a different net economic impact than does the immigration of a long-term welfare recipient.

A nation sets the level of legal immigration through quotas and special provisions. In effect, it also sets the size of illegal immigration through how effectively it secures its borders and enforces its immigration laws. This chapter's Last Word examines the surprising recent decline in illegal immigration into the United States and the various causes of that decline, including increased enforcement activities.

SUMMARY

LO23.1 Describe the extent of legal and illegal immigration into the United States.

Legal immigrants may be either permanent immigrants (green card holders) or temporary immigrants who are legally in the country until a specific date. The United States admitted 1,062,040 legal permanent residents in 2011. About 55 percent of these immigrants were women; 45 percent were men. Roughly 57 percent of legal immigrants were married. The largest number of legal immigrants (143,446) was from Mexico, but Mexicans made up only 14 percent of the total legal immigration in 2011. The vast majority of legal immigrants gain that status because of family ties with U.S. citizens and other legal residents.

Illegal immigrants (also called unauthorized immigrants, illegal aliens, or undocumented workers) are people who enter the country unlawfully or overstay their prescribed exit dates. An estimated 250,000 illegal immigrants enter the United States each year, adding to the 11.1 million illegal immigrants already in the United States. The vast majority of illegal immigrants come from Mexico.

LO23.2 Discuss why economists view economic immigration as a personal human capital investment.

An economic migrant's decision to move to another country can be viewed as an investment, in which present sacrifices (explicit and implicit costs) are incurred to obtain larger lifetime gains (higher earnings). Other things equal, the shorter the distance of the move and the younger the potential economic migrant, the more likely the person will move to the destination country.

LO23.3 Explain how immigration affects average wages, resource allocation, domestic output, and group income shares.

The simple immigration model suggests that, for a high-wage country, the movement of migrants from a low-wage country (*a*) increases domestic output (= domestic income), (*b*) reduces the average wage rate, (*c*) reduces the total wage income of native-born workers, and (*d*) increases business income. The opposite effects occur in the low-wage country. Because the domestic output gains in the high-wage country exceed the domestic output losses in the low-wage country, labor resources are more efficiently allocated globally and world output rises.

The outcomes of immigration predicted by the simple immigration model become more complicated when considering (*a*) the costs of moving, (*b*) the possibility of remittances and backflows, (*c*) complementary rather than substitute labor, (*d*) impacts on investment, (*e*) the levels of unemployment in each country, and (*f*) the fiscal impact on the taxpayers of each country.

LO23.4 Relate how illegal immigration affects employment and wages in low-wage labor markets and impacts state and local budgets.

Legal U.S. residents who have less than a high school education seem to bear the brunt of the wage impact of immigration, although some high-educated workers are also affected. Immigration has little discernable effect on the overall average wage rate in the U.S. economy, with estimates ranging from minus 3 percent to plus 2 percent.

Illegal workers in the United States reduce wage rates in narrowly defined low-wage labor markets, but they do not reduce native-born employment by the full extent of the employment of the illegal workers. American workers who are complementary to illegal immigrant labor may experience an increase in the demand for their services and wages because of the illegal immigrants.

LO23.5 Demonstrate how economics can inform current immigration discussions and attempts to reform immigration laws.

Illegal workers may increase the overall rate of return on capital, thus promoting greater national investment. However, large numbers of illegal workers in specific industries may reduce the incentive for those industries to mechanize. A legitimate concern is that illegal workers and their families impose greater fiscal costs on state and local governments than they contribute in tax revenues to those jurisdictions.

TERMS AND CONCEPTS

economic immigrants
legal immigrants
illegal immigrants
H1-B provision
human capital
beaten paths
backflows
skill transferability
self-selection
efficiency gains from migration
brain drains
remittances
complementary resources
substitute resources
negative self-selection
compensating wage differential

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Which of the following statements are true? Which are false? Explain why the false statements are untrue. **LO23.1**
   a. More immigrants arrive to the United States each year illegally than legally.
   b. The majority of legal immigrants are men.
   c. Over half the new legal immigrants to the United States each year are from Mexico.
   d. Most legal immigrants to the United States gain their legal status through employment-based preferences.
2. In what respect is the economic decision to move across international borders an investment decision? Why do economic migrants move to some countries but not to others? Cite an example of an explicit cost of moving; an implicit cost of moving. How do distance and age affect the migration decision? How does the presence of a large number of previous movers to a country affect the projected costs and benefits of subsequent movers? **LO23.2**
3. Suppose that the projected lifetime earnings gains from migration exceed the costs of moving. Explain how the decision to move might be reversed when a person considers present value. **LO23.2**
4. How might the output and income gains from immigration shown by the simple immigration model be affected by (*a*) unemployment in the originating nation, (*b*) remittances by immigrants to the home country, and (*c*) backflows of migrants to the home country? **LO23.3**
5. Suppose initially that immigrant labor and native-born labor are complementary resources. Explain how a substantial immigration might change the demand for native-born workers, altering their wages. (Review the relevant portion of Chapter 14 if necessary to help answer this question.) Next, suppose that new immigrant labor and previous immigrant labor (not native-born) are substitute resources. Explain how a substantial immigration of new workers might affect the demand for previous immigrants, altering their wages. **LO23.3**
6. What is a "brain drain" as it relates to international migration? If emigrants are highly educated and received greatly subsidized education in the home country, is there any justification for that country to levy a "brain drain" tax on them? Do you see any problems with this idea? **LO23.3**
7. In July 2007 *The Wall Street Journal* (*WSJ*) reported that a growing shortage of skilled labor in Eastern European countries such as Slovakia was driving up wages in key industries and reducing business income. The reason for the shortages was a large migration of skilled Eastern European workers to Western European countries. Use the simple immigration model to demonstrate the key elements of the *WSJ* story as just described. **LO23.3**
8. Why is each of these statements somewhat misleading? (*a*) "Illegal immigrants take only jobs that no American wants." (*b*) "Deporting 100,000 illegal immigrants would create 100,000 job openings for Americans." **LO23.4**

9. Why are so many state and local governments greatly concerned about the federal government's allegedly lax enforcement of the immigration laws and congressional proposals to grant legal status (amnesty) to the 11.1 million illegal immigrants in the United States? How might an amnesty program affect the flow of future border crossings? **LO23.5**
10. If someone favors the free movement of labor within the United States, is it inconsistent for that person to also favor restrictions on the international movement of labor? Why or why not? **LO23.5**
11. **LAST WORD** What was the single most important reason for the decline in illegal immigration between 2007 and 2009? What other factors were at play? Going forward, why might people think it more likely that illegal immigration will fall rather than rise?

REVIEW QUESTIONS

1. Each year, the number of legal immigrants to the United States is __________ the number of illegal immigrants. **LO23.1**
   a. Less than.
   b. Equal to.
   c. Greater than.
   d. Less than, but only in *most* years, not every year.
2. The primary reason people immigrate to the United States is: **LO23.2**
   a. To escape political or religious oppression back home.
   b. To reunite with family members.
   c. To improve earnings and living standards.
   d. None of the above.
3. True or False. Because older adults have more human capital, they are more likely to migrate to another country than younger adults. **LO23.2**
4. Use the accompanying tables for Neon and Zeon to answer the questions that follow. Assume that the wage rate shown equals hourly output and income, and that the accumulated output and income are the sum of the marginal revenue products (MRPs) of each worker. **LO23.3**

Neon

Workers	Wage Rate = MRP	Domestic Output and Income
1	$21	$ 21
2	19	40 (= 21 + 19)
3	17	57 (= 21 + 19 + 17)
4	15	72
5	13	85
6	11	96
7	9	105

Zeon

Workers	Wage Rate = MRP	Domestic Output and Income
1	$15	$15
2	13	28 (= 15 + 13)
3	11	39 (= 15 + 13 + 11)
4	9	48
5	7	55
6	5	60
7	3	63

   a. Which country has the greater stock of capital and technological prowess? How can you tell?
   b. Suppose the equilibrium wage rate is $19 in Neon and $7 in Zeon. What is the domestic output (= domestic income) in the two countries?
   c. Assuming zero migration costs and initial wage rates of $19 in Neon and $7 in Zeon, how many workers will move to Neon? Why will not more than that number of workers move to Neon?
   d. After the move of workers, what will the equilibrium wage rate be in each country? What will the domestic output be after the migration? What is the amount of the combined gain in domestic output produced by the migration? Which country will gain output; which will lose output? How will the income of native-born workers be affected in each country?
5. Migration between North Korea and South Korea has been prohibited since the end of the Korean War in 1953. South Korea is now much richer than North Korea and has a much higher marginal product of labor and a much higher wage rate than North Korea. If workers could migrate from North Korea to South Korea, we would expect: **LO23.3**
   a. Output to fall in South Korea but rise in North Korea.
   b. Output to rise in each country.
   c. Total combined output in the two countries to fall.
   d. Total combined output in the two countries to rise.
6. True or False. Research indicates that immigration causes large decreases in the average American wage. **LO23.3**
7. True or False. The MB = MC level of immigration is likely to be achieved if we simply let in every person who wishes to immigrate to the United States. **LO23.5**

PROBLEMS

1. Mexico has daily (rather than hourly) minimum wage laws. In 2013, the daily minimum wage in Mexico was about 60 pesos per day, and the exchange rate between Mexican pesos and U.S. dollars was about 12 pesos per dollar. **LO23.3**
   a. In 2013, what was the Mexican minimum daily wage in terms of dollars?
   b. Given that Mexican employees typically work 8-hour days, about how much *per hour* is the Mexican minimum wage in terms of dollars?
   c. In 2013, the federal minimum wage in the United States was $7.25 per hour. How many times larger was the hourly U.S. federal minimum wage than the hourly Mexican minimum wage?
   d. If unskilled workers have a tendency to migrate to where they can obtain the highest compensation for their labor, which country is more likely to be receiving low-skilled immigrants?
2. Differences in productivity are usually the major force behind differences in wages and unit labor costs. Suppose that a single unskilled worker at a pottery factory in Mexico can produce 1 mug per hour. By comparison, suppose that a single unskilled worker at a pottery factory in the United States can produce 14 mugs per hour because more and better machinery generates higher labor productivity. The Mexican mugs and the American mugs are identical in quality and durability and sell for the same price. **LO23.3**
   a. If unskilled pottery workers are paid the local minimum wage in both countries, how much is the labor cost *per mug* for mugs produced in Mexico? For mugs produced in the United States? (Use the minimum wages from problem 1 and make all calculations in dollars.)
   b. With regard to mug production, how much higher are labor costs *per hour* in the United States?
   c. With regard to mug production, how much higher are labor costs *per unit* in Mexico?
   d. Do higher labor costs per hour always imply higher labor costs per unit?
   e. If firms with lower labor costs per unit expand, while those with higher labor costs per unit contract, in which country will mug-making firms be increasing in size and hiring more employees? If unskilled pottery workers relocate to where they can find jobs, to which country will they be moving?
3. There is evidence that, other things equal, a 10 percent increase in the number of workers having a particular skill level leads to about a 4 percent decline in wages for workers with that skill level. In addition, this 10-to-4 ratio appears to hold true whether the increase in labor supply is caused by domestic changes in labor supply or by an influx of foreign immigrants. **LO23.3**
   a. Suppose that 42,000 computer programmers work in Silicon Valley. If the number of computer programmers in Silicon Valley increases by 1,260 because of a change in U.S. immigration laws, by how many percentage points would you expect the wage of computer programmers to fall in Silicon Valley?
   b. Suppose that 8,000 full-time cooks work in restaurants in the Denver area. If Denver becomes popular both with U.S. citizens and with foreigners such that 400 full-time cooks move to Denver from other parts of the United States while 80 full-time cooks move to Denver from other countries, by how much would you expect the wages of full-time cooks to fall in Denver?
4. In 2008, an estimated 7.8 million Mexican-born immigrants were employed in the United States. **LO23.3**
   a. If 60 percent of the Mexican-born immigrants remitted money to family members in Mexico in 2008, and if they each sent $100 per month, how much money did they remit in total?
   b. If, instead, 100 percent of the Mexican-born immigrants remitted money to family members in Mexico in 2008, and if they each sent $250 per month, how much money did they remit in total?
   c. The actual amount remitted to Mexico in 2008 by Mexican-born immigrants living in the United States was about $22 billion. If we assume that 75 percent of the Mexican-born immigrants remitted money to Mexico that year and if we further assume that each of those immigrants remitted an equal amount each month, how much per month did each of those immigrants have to remit to total $22 billion for the year?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **SEVEN**

GDP, GROWTH, AND INSTABILITY

CHAPTER **24**

An Introduction to Macroeconomics

Learning Objectives

LO24.1 Explain why economists focus on GDP, inflation, and unemployment when assessing the health of an entire economy.

LO24.2 Discuss why sustained increases in living standards are a historically recent phenomenon.

LO24.3 Identify why saving and investment are key factors in promoting rising living standards.

LO24.4 Describe why economists believe that "shocks" and "sticky prices" are responsible for short-run fluctuations in output and employment.

LO24.5 Characterize the degree to which various prices in the economy are sticky.

LO24.6 Explain why the greater flexibility of prices as time passes causes economists to utilize different macroeconomic models for different time horizons.

Macroeonomics focuses its attention on national economies while seeking answers to the largest of economic questions. For instance: Why are some countries really rich while others are really poor? Why do some countries enjoy sustained, long-run increases in living standards, while other countries simply stagnate? Why do all countries—even the richest—go through alternating boom and bust periods? And is there anything that governments can do to improve living standards or fight recessions?

This chapter provides an overview of the data that macroeconomists use to measure the status and growth of an entire economy as well as a preview of the models that they use to help explain both long-run growth and short-run fluctuations. Because it is an overview chapter, it raises many unanswered questions. Subsequent chapters will explain these topics in much greater detail.

Performance and Policy

LO24.1 Explain why economists focus on GDP, inflation, and unemployment when assessing the health of an entire economy.

As you know from Chapter 1, macroeconomics studies the behavior of the economy as a whole. It is primarily concerned with two topics: long-run economic growth and the short-run fluctuations in output and employment that are often referred to as the **business cycle.** These phenomena are closely related because they happen simultaneously. Economies show a distinct growth trend that leads to higher output and higher standards of living in the long run, but in the short run there is considerable variability. Sometimes growth proceeds more rapidly and sometimes it proceeds more slowly. It may even turn negative for a while so that output and living standards actually decline, a situation referred to as a **recession.** That is precisely what happened in late 2007 and continued through 2008 and into 2009. The economy experienced what has come to be called the Great Recession.

To understand how economies operate and how their performance might be improved, economists collect and analyze economic data. An almost infinite number of data items can be looked at, including the amount of new construction taking place each month, how many ships laden with cargo are arriving at our ports each year, and how many new inventions have been patented in the last few weeks. That being said, macroeconomists tend to focus on just a few statistics when trying to assess the health and development of an economy. Chief among these are real GDP, unemployment, and inflation.

- **Real GDP,** or **real gross domestic product,** measures the value of final goods and services produced within the borders of a country during a specific period of time, typically a year. This statistic is very useful because it can tell us whether an economy's output is growing. For instance, if U.S. real GDP in one year is larger than in the previous year, we know that U.S. output increased from the first year to the next. To determine real GDP, government statisticians first calculate **nominal GDP,** which totals the dollar value of all goods and services produced within the borders of a country using *their current prices during the year that they were produced.* But because nominal GDP uses the prices in place in the year the output was produced, it suffers from a major problem: It can increase from one year to the next even if there is no increase in output. To see how, consider a commercial blacksmith who produced 10 iron spiral staircases last year and 10 identical staircases this year. Clearly, the blacksmith's output did not change. But if the price of each staircase rose from \$10,000 last year to \$20,000 this year, nominal GDP increased from \$100,000 (= 10 × \$10,000) to \$200,000 (= 10 × \$20,000). Without knowing about the price increase, we might unwisely conclude that the output of staircases increased from 10 to 20. Real GDP statistically eliminates these kinds of price changes. As a result, we can compare real GDP numbers from one year to the next and really know if there is a change in output (rather than prices). Because more output means greater consumption possibilities—including not only the chance to consume more fun things such as movies, vacations, and video games, but also more serious things like better health care and safer roads—economists and policymakers are deeply committed to encouraging a large and growing real GDP.
- **Unemployment** is the state a person is in if he or she cannot get a job despite being willing to work and actively seeking work. High rates of unemployment are undesirable because they indicate that a nation is not using a large portion of its most important resource—the talents and skills of its people. Unemployment is a waste because we must count as a loss all the goods and services that unemployed workers could have produced if they had been working. Researchers have also drawn links between higher rates of unemployment and major social problems like higher crime rates and greater political unrest as well as higher rates of depression, heart disease, and other illnesses among unemployed individuals.
- **Inflation** is an increase in the overall level of prices. As an example, consider all the goods and services bought by a typical family over the course of one year. If the economy is experiencing inflation, it will cost the family more money to buy those goods and services this year than it cost to buy them last year. This can be

troublesome for several reasons. First, if the family's income does not rise as fast as the prices of the goods and services that it consumes, it won't be able to purchase as much as it used to and its standard of living will fall. Along the same lines, a surprise jump in inflation reduces the purchasing power of people's savings. Savings that they believed would be able to buy them a specific amount of goods and services will turn out to buy them less than they expected due to the higher-than-expected prices.

Because these statistics are the standards by which economists keep track of long-run growth and short-run fluctuations, we will spend a substantial amount of time in the next few chapters examining how these statistics are computed, how well they are able to capture the well-being of actual people, and how they vary both across countries and over time. Once they are understood, we will build on them in subsequent chapters by developing macroeconomic models of both long-run growth and short-run fluctuations. These will help us understand how policymakers attempt to maximize growth while minimizing unemployment and inflation.

Macroeconomic models also clarify many important questions about the powers and limits of government economic policy. These include:

- Can governments promote long-run economic growth?
- Can they reduce the severity of recessions by smoothing out short-run fluctuations?
- Are certain government policy tools such as manipulating interest rates (monetary policy) more effective at mitigating short-run fluctuations than other government policy tools such as changes in tax rates or levels of government spending (fiscal policy)?
- Is there a trade-off between lower rates of unemployment and higher rates of inflation?
- Does government policy work best when it is announced in advance or when it is a surprise?

The answers to these questions are of crucial importance because of the vast differences in economic performance experienced by national economies at different times. For instance, the amount of output generated by the U.S. economy grew at an average rate of 2.7 percent per year between 1995 and the start of the recession of 2007–2009, while the amount of output generated by the Japanese economy grew at an average rate of only 1.0 percent per year over the same time period. In 2008 and 2009, however, the U.S. economy lost 8 million jobs, and the unemployment rate rose from 4.6 percent to as high as 10.1 percent. A couple of years later, in 2011, the unemployment rate was 9.1 percent in the United States, 17.9 percent in Greece, 3.5 percent in South Korea, 9.3 percent in France, and 40.6 percent in Haiti. At the same time, the inflation rate was 3.0 percent in the United States, 1.3 percent in Norway, 14.0 percent in Kenya, 21.0 percent in Argentina, and 3.4 percent in Mexico.

Our models will help us understand why such large differences in rates of growth, unemployment, and inflation exist among countries and why those rates can change so substantially from one period to another. These models also will provide significant insights on how government policies can influence rates of growth, unemployment, and inflation.

QUICK REVIEW 24.1

- Macroeconomics studies long-run economic growth and short-run economic fluctuations.
- Macroeconomists focus their attention on three key economic statistics: real GDP, unemployment, and inflation.
- Macroeconomic models help to clarify many important questions about government economic policy.

The Miracle of Modern Economic Growth

LO24.2 Discuss why sustained increases in living standards are a historically recent phenomenon.

Rapid and sustained economic growth is a modern phenomenon. Before the Industrial Revolution began in the late 1700s in England, standards of living showed virtually no growth over hundreds or even thousands of years. For instance, the standard of living of the average Roman peasant was virtually the same at the start of the Roman Empire around the year 500 B.C. as it was at the end of the Roman Empire 1,000 years later. Similarly, historians and archaeologists have estimated that the standard of living enjoyed by the average Chinese peasant was essentially the same in the year A.D. 1800 as it was in the year A.D. 100.

That is not to say that the Roman and Chinese economies did not expand over time. They did. In fact, their total outputs of goods and services increased many times over. The problem was that as they did, their populations went up by similar proportions so that the amount of output *per person* remained virtually unchanged.

This historical pattern continued until the start of the Industrial Revolution, which ushered in not only factory production and automation but also massive increases in research and development so that new and better technologies were constantly being invented. The result was that output began to grow faster than the population. This meant that living standards began to rise as the amount of output *per person* increased.

Not all countries experienced this phenomenon, but those that did were said to be experiencing **modern economic growth** (in which output per person rises) as compared with earlier times in which output (but not output per person) increased. Under modern economic growth, the annual increase in output per person is often not large, perhaps 2 percent per year in countries such as England that were the first to industrialize. But when compounded over time, an annual growth rate of 2 percent adds up very rapidly. Indeed, it implies that the standard of living will double every 35 years. So if the average citizen of a country enjoying 2 percent growth begins this year with an income of $10,000, in 35 years that person will have an income of $20,000. And 35 years after that there will be another doubling so that her income in 70 years will be $40,000. And 35 years after that, the average citizen's income will double again to $80,000. Such high rates of growth are amazing when compared to the period before modern economic growth when standards of living remained unchanged century after century.

The vast differences in living standards seen today between rich and poor countries are almost entirely the result of the fact that only some countries have experienced modern economic growth. Indeed, before the start of the Industrial Revolution in the late 1700s, living standards around the world were very similar, so much so that the average standard of living in the richest parts of the world was at most only two or three times higher than the standard of living in the poorest parts of the world. By contrast, the citizens of the richest nations today have material standards of living that are on average more than 50 times higher than those experienced by citizens of the poorest nations, as can be seen by the GDP per person data for the year 2011 given in Global Perspective 24.1.

Global Perspective 24.1 facilitates international comparisons of living standards by making three adjustments to each country's GDP. First, it converts each country's GDP from its own currency into U.S. dollars so that there is no confusion about the values of different currencies. Second, it divides each country's GDP measured in dollars by the size of its population. The resulting number, *GDP per person*, is the average amount of output each person in each country could have if each country's total output were divided equally among its citizens. It is a measure of each country's average standard of living. Third, the table uses a method called *purchasing power parity* to adjust for the fact that prices are much lower in some countries than others. By making this adjustment, we can trust that $1 of GDP per person in the United States represents about the same quantity of goods and services as $1 of GDP per person in any of the other countries. The resulting numbers—GDP per person adjusted for purchasing power parity—are presented in Global Perspective 24.1.

GLOBAL PERSPECTIVE 24.1

GDP per Person, Selected Countries

Country	GDP per Person, 2011 (U.S. dollars based on purchasing power parity)
Canada	$50,496
United States	48,328
Japan	45,870
France	44,007
United Kingdom	38,811
South Korea	31,700
Saudi Arabia	21,196
Russia	12,993
Mexico	10,146
China	5,417
North Korea	1,800
India	1,514
Zimbabwe	752
Tanzania	566
Burundi	275

Source: International Monetary Fund, **www.imf.org**, for all countries except for North Korea, the estimate for which is from the *CIA World Factbook*, **www.cia.gov**.

QUICK REVIEW 24.2

- Before the Industrial Revolution, living standards did not show any sustained increases over time because any increase in output tended to be offset by an equally large increase in population.
- Since the Industrial Revolution, many nations have experienced *modern economic growth* in which output grows faster than population—so that living standards rise over time.

Saving, Investment, and Choosing between Present and Future Consumption

LO24.3 Identify why saving and investment are key factors in promoting rising living standards.

At the heart of economic growth is the principle that to raise living standards over time, an economy must devote at least some fraction of its current output to increasing future output. As implied in Chapter 1, this process requires flows of both saving and investment, which we will define and discuss before returning to why they are so important for economic growth.

- **Saving** occurs when current consumption is less than current output (or when current spending is less than current income).
- **Investment** happens when resources are devoted to increasing future output—for instance by building a new research facility in which scientists invent the next generation of fuel-efficient automobiles or by constructing a modern, super-efficient factory. (A caution: In economics, the term "investment" differs from common usage. To understand why, be sure to read the Consider This box.)

When thinking about why saving and investment are so important for economic growth, the key point is that the amount of investment is ultimately limited by the amount of saving. The only way that more output can be directed at investment activities is if saving increases. But that, in turn, implies that individuals and society as a whole must make trade-offs between current and future consumption. This is true because the only way to pay for more investment—and the higher levels of future consumption that more investment can generate—is to increase present saving. But increased saving can only come at the price of reduced current consumption. Individuals and society as a whole must therefore wrestle with a choice between present consumption and future consumption. They must decide how to balance the reductions in current consumption required to fund current investment against the increases in future consumption that the added current investment will make possible.

CONSIDER THIS . . .

Economic versus Financial Investment

Economics students often are confused by how the word "investment" is used in economics. This is understandable, because economists draw a distinction between "financial investment" and "economic investment."

Financial investment captures what ordinary people mean when they say investment, namely, the purchase of assets like stocks, bonds, and real estate in the hope of reaping a financial gain. Anything of monetary value is an asset and, in everyday usage, people purchase—or "invest" in—assets hoping to receive a financial gain, either by eventually selling them at higher prices than they paid for them or by receiving a stream of payments from others who are allowed to use the asset. By contrast, when economists say "investment," they are referring to **economic investment,** which relates to the creation and expansion of business enterprises. Specifically, economic investment only includes spending on the production and accumulation of newly created capital goods such as machinery, tools, factories, and warehouses. For example, economic investment will occur when the airplane shown in the accompanying photo is purchased by a commercial airline.

For economists, purely financial transactions, such as swapping cash for a stock or a bond, are not "investment." Neither are the purchases of factories or apartment buildings built in previous years. These transactions simply transfer the ownership of financial assets or existing real assets from one party to another. They do not purchase newly created capital goods. As such, they are great examples of financial investment, but not of economic investment. So now that you know the difference, remember that purely financial transactions, like buying Google stock or a five-year-old factory, are indeed referred to as "investment"—except in economics!

Banks and Other Financial Institutions

Households are the principal source of savings. But businesses are the main economic investors. So how does the pool of savings generated by households when they spend less than they consume get transferred to businesses so that they can purchase newly created capital goods? The answer is through banks and other financial institutions such as mutual funds, pension plans, and insurance companies. These institutions collect the savings of households, rewarding savers with interest and dividends and sometimes capital gains (increases in asset values). The banks and other financial institutions then lend the funds to businesses, which invest in equipment, factories, and other capital goods.

Macroeconomics devotes considerable attention to money, banking, and financial institutions because a well-functioning financial system helps to promote economic growth and stability by encouraging saving and by properly directing that saving into the most productive possible investments. In contrast, a poorly functioning financial system can create serious problems for an economy.

QUICK REVIEW 24.3

- An economy can only grow if it invests, and it can only invest if it saves some of its current output. Thus, saving is crucial to increasing investment and, consequently, future output.
- Banks and other financial institutions channel household savings toward businesses, which invest in equipment, factories, and other capital goods.

Uncertainty, Expectations, and Shocks

LO24.4 Describe why economists believe that "shocks" and "sticky prices" are responsible for short-run fluctuations in output and employment.

Decisions about savings and investment are complicated by the fact that the future is uncertain. Investment projects sometimes produce disappointing results or even fail totally. As a result, firms spend considerable time trying to predict future trends so that they can, hopefully, invest only in projects that are likely to succeed. This implies that macroeconomics has to take into account **expectations** about the future.

The Importance of Expectations and Shocks

Expectations are hugely important for two reasons. The more obvious reason involves the effect that changing expectations have on current behavior. If firms grow more pessimistic about the future returns that are likely to come from current investments, they are going to invest less today than they would if they were more optimistic. Expectations therefore have a large effect on economic growth since increased pessimism will lead to less current investment and, subsequently, less future consumption.

The less-obvious reason that expectations are so important has to do with what happens when expectations are unmet. Firms are often forced to cope with **shocks**—situations in which they were expecting one thing to happen but then something else happened. For instance, consider a situation in which a firm decides to build a high-speed railroad that will shuttle passengers between Los Angeles and Las Vegas. The firm expects it to be very popular and make a handsome profit. But if it unexpectedly turns out to be unpopular and loses money, the railroad must figure out how to respond. Should the railroad go out of business completely? Should it attempt to see if it can turn a profit by hauling cargo instead of passengers? Is there a possibility that the venture might succeed if the firm borrows $30 million from a bank to pay for a massive advertising campaign? These sorts of decisions are necessitated by the shock and surprise of having to deal with an unexpected situation.

Economies are exposed to both demand shocks and supply shocks. **Demand shocks** are unexpected changes in the demand for goods and services. **Supply shocks** are unexpected changes in the supply of goods and services. Note that the word *shock* only reveals that something unexpected has happened. It does not tell us whether what has happened is unexpectedly good or unexpectedly bad. To clarify this, economists use more specific terms. For instance, a *positive demand shock* refers to a situation in which demand turns out to be higher than expected, while a *negative demand shock* refers to a situation in which demand turns out to be lower than expected.

Demand Shocks and Sticky Prices

Economists believe that most short-run fluctuations in GDP and the business cycle are the result of demand shocks. Supply shocks do happen in some cases and are very important when they do occur. But we will focus most of our attention in this chapter and subsequent chapters on demand shocks, how they affect the economy, and how government policy may be able to help the economy adjust to them.

But why are demand shocks such a big problem? Why would we have to consider calling in the government to help deal with them? And why can't firms deal with demand shocks on their own?

The answer to these questions is that the prices of many goods and services are inflexible (slow to change, or "sticky") in the short run. As we will explain, this implies that price changes do not quickly equalize the quantities demanded of such goods and services with their respective quantities supplied. Instead, because prices are inflexible, the economy is forced to respond in the short run to demand shocks primarily through changes in output and employment rather than through changes in prices.

Example: A Single Firm Dealing with Demand Shocks and Sticky Prices

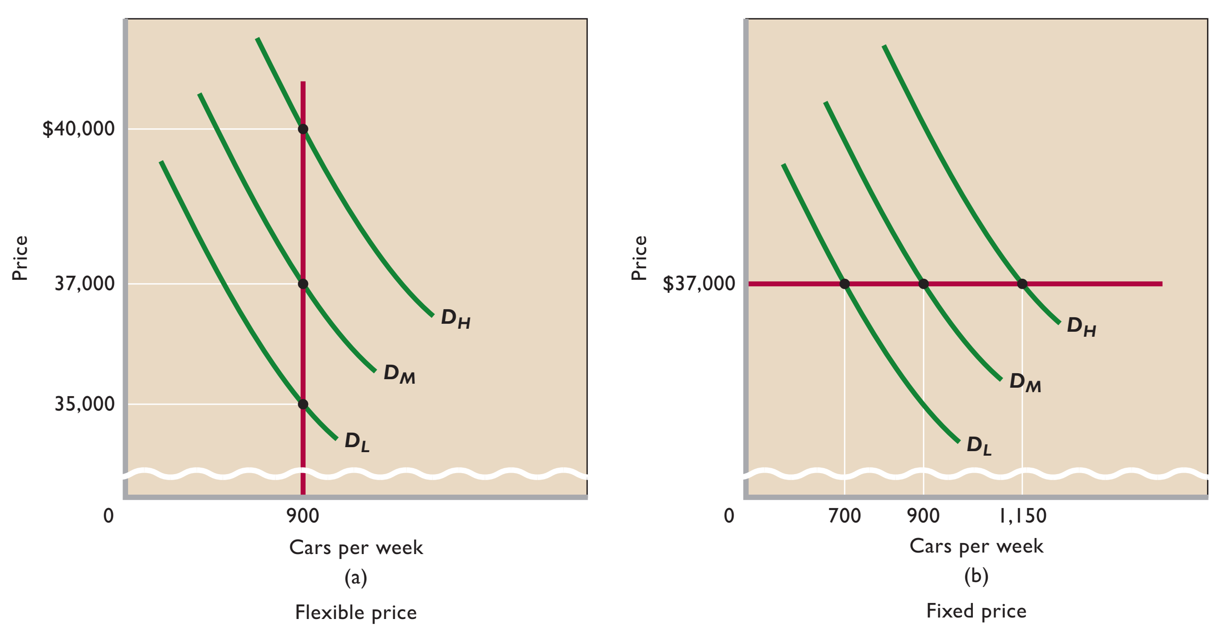
Although an economy as a whole is vastly more complex than a single firm, an analogy that uses a single car factory will be helpful in explaining why demand shocks and inflexible prices are so important to understanding most of the short-run fluctuations that affect the entire economy. Consider a car manufacturing company named Buzzer Auto. Like most companies, Buzzer Auto is in business to try to make a profit. Part of turning a profit involves trying to develop accurate expectations about future market conditions. Consequently, Buzzer constantly does market research to estimate future demand conditions so that it will, hopefully, only build cars that people are going to want to buy.

Setting Expectations After extensive market research, Buzzer concludes that it could earn a modest profit if it builds and staffs an appropriately sized factory to build an environmentally friendly SUV, which it decides to call the Prion. Buzzer's marketing economists collaborate with Buzzer's engineers and conclude that expected profits will be maximized if the firm builds a factory that has an optimal output rate of 900 cars per week. If the factory operates at this rate, it can produce Prions for only $36,500 per vehicle. This is terrific because the firm's estimates for demand indicate that a supply of 900 vehicles per week can be sold at a price of $37,000 per vehicle—meaning that if everything goes according to plan, Buzzer Auto should make an accounting profit of $500 on each Prion that it produces and sells. Expecting these future conditions, Buzzer decides to build the factory, staff it with workers, and begin making the Prion.

Look at Figure 24.1a, which shows the market for Prions when the vertical supply curve for Prions is fixed at the factory's optimal output rate of 900 cars per week. Notice that we have drawn in three possible demand curves. D_L corresponds to low demand for the Prion; D_M corresponds to the medium level of demand that Buzzer's marketing economists are expecting to materialize; and D_H corresponds to high demand for the Prion. Figure 24.1a is consistent with the marketing economists' expectations: if all goes according to plan and the actual demand that materializes is D_M, the equilibrium price will in fact be $37,000 per Prion and the equilibrium quantity demanded will be 900 cars per week. Thus, if all goes according to expectations, the factory will have exactly the right capacity to meet the expected quantity demanded at the sales price of $37,000 per vehicle. In addition, the firm's books will show a profit of $500 per vehicle on each of the 900 vehicles that it builds and expects to sell each week at that price.

Full Employment If There Are No Shocks Here is the key point. If expectations are always fulfilled, Buzzer Auto will never contribute to any of the short-run fluctuations in output and unemployment that affect real-world economies. First, if everything always goes according to plan and Buzzer Auto's expectations always come true, then the factory will always produce and sell at its optimal

FIGURE 24.1 The effect of unexpected changes in demand under flexible and fixed prices. (a) If prices are flexible, then no matter what demand turns out to be, Buzzer Auto can continue to sell its optimal output of 900 cars per week since the equilibrium price will adjust to equalize the quantity demanded with the quantity supplied. (b) By contrast, if Buzzer Auto sticks with a fixed-price policy, then the quantity demanded will vary with the level of demand. At the fixed price of $37,000 per vehicle, the quantity demanded will be 700 cars per week if demand is D_L, 900 cars per week if demand is D_M, and 1,150 cars per week if demand is D_H.



output rate of 900 cars per week. This would mean that it would never experience any fluctuations in output—either in the short run or in the long run. At the same time, since producing a constant output of 900 cars each week will always require the same number of workers, the factory's labor demand and employment should never vary. So if everything always goes according to plan, Buzzer Auto will never have any effect on unemployment because it will always hire a constant number of workers.

These facts imply that the short-run fluctuations in output and unemployment that we do see in the real world must be the result of shocks and things *not* going according to plan. In particular, business cycle fluctuations typically arise because the actual demand that materializes ends up being either lower or higher than what people were expecting. When this occurs, some adjustments will be necessary to bring quantity demanded and quantity supplied back into alignment. As we are about to explain, the nature of these adjustments varies hugely depending on whether prices are flexible or inflexible.

Price Changes If There Are Demand Shocks and Flexible Prices Figure 24.1a illustrates the case of adjusting to unexpected changes in demand *when prices are flexible*. Here, if demand is unexpectedly low at D_L, the market price can adjust downward to $35,000 per vehicle so that the quantity demanded at that price will still be equal to the factory's optimal output rate of 900 cars per week. On the other hand, if demand is unexpectedly high at D_H, the market price can adjust upward to $40,000 per vehicle so that the quantity demanded will still be equal to the factory's optimal output rate of 900 cars per week. These adjustments imply that *if* the price of Prions is free to quickly adjust to new equilibrium levels in response to unexpected changes in demand, the factory could always operate at its optimal output rate of 900 cars per week. Only the amount of profit or loss will vary with demand.

Applying this logic to the economy as a whole, *if* the prices of goods and services could always adjust quickly to unexpected changes in demand, then the economy could always produce at its optimal capacity since prices would adjust to ensure that the quantity demanded of each good and service would always equal the quantity supplied. Simply put, if prices were fully flexible, there would be no short-run fluctuations in output. Production levels would remain constant and unemployment levels would not change because firms would always need the same number of workers to produce the same amount of output.

Output Changes If There Are Demand Shocks and Sticky Prices In reality, many prices in the economy are inflexible and are not able to change rapidly when demand changes unexpectedly. Consider the extreme case shown in Figure 24.1b, in which the price of Prions is totally inflexible, fixed at $37,000 per Prion. Here, if demand unexpectedly falls from D_M to D_L, the quantity demanded at the fixed price of $37,000 will only be 700 cars per week, which is 200 cars fewer than the factory's optimal output of 900 cars per week. On the other hand, if demand is unexpectedly high at D_H, the quantity demanded at the fixed price of $37,000 will be 1,150 cars per week, which is 250 cars more than the factory's optimal output of 900 cars per week.

One way for companies to deal with these unexpected shifts in quantity demanded would be to try to adjust the factory's output to match them. That is, during weeks of low demand, Buzzer Auto could attempt to produce only 700 Prions, while during weeks of high demand it could try to produce 1,150 Prions. But this sort of flexible output strategy is very expensive because factories operate at their lowest costs when they are producing constantly at their optimal output levels; operating at either a higher or a lower production rate results in higher per-unit production costs.[1]

Knowing this, manufacturing firms typically attempt to deal with unexpected changes in demand by maintaining an inventory. An **inventory** is a store of output that has been produced but not yet sold. Inventories are useful because they can be allowed to grow or decline in periods when demand is unexpectedly low or high—thereby allowing production to proceed smoothly even when demand is variable. In our example, Buzzer Auto would maintain an inventory of unsold Prions. In weeks when demand is unexpectedly low, the inventory will increase by 200 Prions as the quantity demanded falls 200 vehicles short of the factory's optimal output. By contrast, during weeks when demand is unexpectedly high, the inventory will decrease as the quantity demanded exceeds the factory's optimal output by 250 cars. By allowing inventory levels to fluctuate with these unexpected shifts in demand, Buzzer Auto can respond by adjusting inventory levels rather than output levels. In addition, with any luck, the overall inventory level will stay roughly constant over time as unexpected increases and decreases in demand cancel each other out.

But consider what will happen if the firm experiences many successive weeks of unexpectedly low demand. For each such week, the firm's inventory of unsold Prions will increase by 200 cars. The firm's managers will not mind if

[1]If you have studied microeconomics, you will recognize that the firm's optimal output level of 900 cars per week is the level that minimizes the factory's average total cost (ATC) per vehicle of producing the Prion. Producing either more or fewer Prions will result in higher per-vehicle production costs.

this happens for a few weeks, but if it continues for many weeks, then the managers will be forced to cut production because, among other things, there will simply be no place to park so many unsold vehicles. More importantly, holding large numbers of unsold cars in inventory is unprofitable because while costs must be incurred to build an unsold car, an unsold car obviously brings in no revenue. Constantly rising inventories hurt firm profits and the management will want to reduce output if it sees inventories rising week after week due to unexpectedly low demand.

Generalizing from a Single Firm to the Entire Economy

This simplified story about a single car company explains why economists believe that a combination of unexpected changes in demand and inflexible prices are the key to understanding the short-run fluctuations that affect real-world economies. If prices were flexible, then the firm could always operate at the factory's optimal output level because prices would always adjust to ensure that it could sell its optimal output of 900 cars per week no matter what happens to demand. But if prices are inflexible, then an unexpected decline in demand that persists for any length of time will result in increasing inventories that will eventually force the firm's management to cut production to less than the optimal output level of 900 cars per week. When this happens, not only will output fall, but unemployment will also rise. The firm will lay off workers because fewer employees will be needed to produce fewer cars.

Generalizing this story to the economy as a whole, if demand falls off for many goods and services across the entire economy for an extended period of time, then the firms that make those goods and services will be forced to cut production. Manufacturing firms that maintain inventories will do so as they find inventories piling up due to sluggish sales. And services firms will do so as they encounter slow sales for their services. As both manufacturing and service output declines, the economy will recede, with GDP falling and unemployment rising.

On the other hand, if demand is unexpectedly high for a prolonged period of time, the economy will boom and unemployment will fall. In the case of our Prion example, for each week that demand is unexpectedly high, inventories will fall by 250 cars. If this keeps happening week after week, inventories will start to run out and the firm will have to react by increasing production to more than the optimal output rate of 900 cars per week so that orders do not go unfilled. When this happens, GDP will increase as more cars per week are produced and unemployment will fall because the factory will need to hire more workers to produce the larger number of cars.

CONSIDER THIS . . .

The Great Recession

In 2008 and 2009, the United States encountered its worst financial and economic crisis since the Great Depression of the 1930s.

The recession was so severe that it has been dubbed the Great Recession. The recession was triggered by a steep decline in housing prices and a crisis involving mortgage loans and the financial securities built on them. Several key U.S. financial institutions collapsed or nearly failed, and lending markets largely froze. Despite government bailout efforts, the financial crisis eventually spread to the broader economy. Employment fell by 8 million workers between 2007 and the end of 2009, and the unemployment rate rose from 4.6 percent to 10.1 percent over that same period. Economic growth slumped to 0.4 percent in 2008 and to a *negative* 2.4 percent in 2009, compared with the 2.7 percent annual increases occurring between 1995 and 2007.

And this is where Buzzer Auto comes into the picture. The situation in Figure 24.1b, where the price of Buzzer's autos is inflexible, is highly relevant to the Great Recession. Like Buzzer, actual auto producers such as GM, Ford, and Chrysler, as well as thousands of producers of other products across the economy, established their production capacity and set their expectations of product demand on the basis of normal times. But demand for their goods and services fell unexpectedly because of greater consumer difficulty in getting loans, declining consumer confidence, and eventually declining income. The economy's price level (essentially a weighted average of all prices) declined only slightly, and that was after the recession was well underway. Therefore, real output (not prices) took the major brunt of the decline of total demand in the economy. Output dropped, employment plummeted, and unemployment soared.

QUICK REVIEW 24.4

- Economic shocks occur when events unfold in ways that people were not expecting.
- Demand and supply shocks take place when demand or supply ends up being either higher or lower than expected.
- Real-world prices are often inflexible or "sticky" in the short run.
- When prices are sticky, the economy adjusts to demand shocks mostly through changes in output and employment (rather than through changes in prices).

LAST WORD

Debating the Great Recession

Economists Disagreed Vigorously about Both the Causes of the Great Recession and the Best Ways to Speed a Recovery.

The Great Recession of 2007–2009 was the worst economic downturn since the Great Depression of the 1930s. The government intervened massively to help promote recovery, but the recession was long-lasting and the subsequent recovery was the weakest since the Great Depression.

Explanations about what caused the Great Recession differ sharply among economists. Here are two of the more popular hypotheses.

The Minksy Explanation: Euphoric Bubbles Economist Hyman Minksy believed that severe recessions are often preceded by *asset-price bubbles*—periods during which euphoria and debt-fueled speculation cause the price of one or more financial assets to irrationally skyrocket before collapsing down to more realistic levels. Those who apply his ideas to the Great Recession note that easily obtained home-mortgage loans drove a massive bubble in housing prices.

When the bubble eventually collapsed, investors lost trillions of dollars in wealth. As a result, the demand for goods and services fell dramatically and unexpectedly. When combined with sticky prices, that leftward shift in demand forced many companies to reduce output and lay off workers (as in our Buzzer Auto example in this chapter). The weakest firms went bankrupt and had to permanently fire all of their workers.

The Austrian Explanation: Excessively Low Interest Rates Economists of the so-called Austrian School also blame bubbles for severe recessions, but they put the blame for bubbles not on euphoria but on government actions that they say keep interest rates too low. Their contention is that excessively low interest rates induce firms and individuals to borrow excessively. Individuals borrow excessively to fund consumption. Firms borrow excessively for construction and investment. When the bubble pops, society has too many factories (as a result of the massive increase in construction and investment on the part of firms) combined with too little demand (as consumers struggle to repay all the money they borrowed to fund their consumption).

In terms of this chapter's Buzzer Auto example, it would be as though Buzzer borrowed lots of money to build several factories only to discover that demand was much lower than expected because consumers were cutting back on spending in order to repay debt. With demand shifting left and prices sticky, Buzzer and other companies are forced to reduce output and lay off workers. Thus begins the recession.

How Sticky Are Prices?

LO24.5 Characterize the degree to which various prices in the economy are sticky.

We have just shown that **inflexible prices**—or **"sticky prices"** as economists are fond of saying—help to explain how unexpected changes in demand lead to the fluctuations in GDP and employment that occur over the course of the business cycle. Of course, not all prices are sticky. Indeed, the markets for many commodities and raw materials such as corn, oil, and natural gas feature extremely **flexible prices** that react within seconds to changes in supply and demand. By contrast, the prices of most of the final goods and services that people consume are quite sticky, with the average good or service going 4.3 months between price changes. To get a better appreciation for the fact that price stickiness varies greatly by product or service, look at Table 24.1, which gives the average number of months between price

TABLE 24.1 Average Number of Months between Price Changes for Selected Goods and Services

Item	Months
Coin-operated laundry machines	46.4
Newspapers	29.9
Haircuts	25.5
Taxi fare	19.7
Veterinary services	14.9
Magazines	11.2
Computer software	5.5
Beer	4.3
Microwave ovens	3.0
Milk	2.4
Electricity	1.8
Airline tickets	1.0
Gasoline	0.6

Source: Mark Bils and Peter J. Klenow, "Some Evidence on the Importance of Sticky Prices." *Journal of Political Economy,* October 2004, pp. 947–985. Used with permission of The University of Chicago Press via Copyright Clearance Center.

Because economists did not have a consensus about what caused the Great Recession, it should not be surprising that they were also divided over the best policies for fighting the recession and improving upon the sluggish recovery that began in 2009. For simplicity, the wide variety of opinions can be grouped into two broad camps promoting two very different solutions.

The Stimulus Solution The majority of economists argued that the solution to the collapse in demand was to have the government take actions to shift demand curves rightward. For instance, the government could lower interest rates so that consumers and businesses would borrow and spend more. The government could also massively increase its purchases of goods and services so that a rightward shift in the government's demand for output could help to make up for the leftward shift in the private-sector demand for output.

This opinion in favor of *government stimulus* was the most commonly held view among economists and the government did in fact push interest rates very low while also massively increasing government spending.

The Structural Solution A vocal minority of economists rejected the stimulus policies. They argued that the economy required a *structural adjustment*. In their opinion, the bubble period before the recession had seen a major misallocation of resources toward inefficient firms that generated net losses for society (MB < MC). The only way to redirect the resources that those firms were using back toward productive activities would be to let the inefficient firms go bankrupt. The resources would then flow toward efficient firms whose output generated net benefits for society (MB > MC).

Under this way of thinking, government stimulus efforts delayed recovery by keeping many wasteful firms on life support. Those who took that opinion wanted the government to mostly hang back, let inefficient firms go bankrupt, and allow the invisible hand to reallocate resources.

This debate over government stimulus was ongoing and continual during the sluggish recovery from the Great Recession. Those in favor of stimulus argued that the sluggish recovery was the result of too little stimulus. Those against stimulus argued that the sluggish recovery was the result of too much stimulus.

One of your tasks as you work your way through the subsequent chapters will be to understand the nature of this debate and the arguments and evidence on both sides. But don't look for a definitive answer. The complexities of giant national economies are only partly understood and the best policy may turn out to be something unseen by either of the two camps.

changes for various common goods and services. The prices of some products like gasoline and airline tickets change very rapidly—about once a month or even less than once a month. By contrast, haircuts and newspapers average more than two years between price changes. And coin-operated laundry machines average nearly four years between price changes!

An important recent study has found that product prices are particularly sticky in response to widespread macroeconomic and monetary disturbances.[2] In later chapters, we will identify and discuss several factors that cause short-run price stickiness. But to keep the current discussion brief, let's focus on just two factors here. One factor is that companies selling final goods and services know that consumers prefer stable, predictable prices that do not fluctuate rapidly with changes in demand. Consumers would be annoyed if the same bottle of soda or shampoo cost one price one day, a different price the next day, and yet another price a week later. Volatile prices make planning more difficult, and, in addition, consumers who come in to buy the product on a day when the price happens to be high will likely feel that they are being taken advantage of. To avoid this, most firms try to maintain stable prices that do not change very often. Firms do have occasional sales where they lower prices, but on the whole they tend to try to keep prices stable and predictable—the result being price inflexibility.

Another factor that causes sticky prices has to do with the fact that in certain situations, a firm may be afraid that cutting its price may be counterproductive because its rivals might simply match the price cut—a situation often referred to as a "price war." This possibility is common among firms that only have one or two

[2]Jean Boivin, Marc P. Giannoni, and Illian Mihov, "Sticky Prices and Monetary Policy: Evidence from Disaggregated US Data," *American Economic Review*, March 2009, pp. 350–384.

major rivals. Consider Coca-Cola and Pepsi. If Coca-Cola faces unexpectedly low demand for its product, it might be tempted to reduce its price in the hope that it can steal business away from Pepsi. But such a strategy would only work if Pepsi left its price alone when Coca-Cola cut its price. That, of course, is not likely. If Coca-Cola cuts its price, Pepsi will very likely cut its price in retaliation, doing its best to make sure that Coca-Cola doesn't steal away any of its customers. Thus, if Pepsi retaliates, Coca-Cola will only be made worse off by its decision to cut its price: It will not pick up much more business (because Pepsi also cut its price) and it will also be receiving less money for each bottle of Coke that it sells (because it lowered its own price.) Thus, firms that have to deal with the possibility of price wars often have sticky prices.

Categorizing Macroeconomic Models Using Price Stickiness

LO24.6 Explain why the greater flexibility of prices as time passes causes economists to utilize different macroeconomic models for different time horizons.

We have now demonstrated why price stickiness is believed to have such a large role in short-run economic fluctuations. It should be noted, however, that price stickiness moderates over time. This is true because firms that choose to use a fixed-price policy in the short run do not have to stick with that policy permanently. In particular, if unexpected changes in demand begin to look permanent, many firms will allow their prices to change so that price changes (in addition to quantity changes) can help to equalize quantities supplied with quantities demanded.

For this reason, economists speak of "sticky prices" rather than "stuck prices." Only in the very short run are prices totally inflexible. As time passes and prices are revised, the world looks much more like Figure 24.1a, in which prices are fully flexible, rather than Figure 24.1b, in which prices are totally inflexible. Indeed, the totally inflexible case shown in Figure 24.1b can be thought of as the extremely short-run response to an unexpected change in demand, while the fully flexible case shown in Figure 24.1a can be thought of as a longer-run response to an unexpected change in demand. In terms of time durations, the extreme short run can be thought of as the first few weeks and months after a demand shock, while the long run can be thought of as extending from many months to several years after a demand shock happens.

This realization is very useful in categorizing and understanding the differences between the various macroeconomic models that we will be presenting in subsequent chapters. For instance, the aggregate expenditures model presented in Chapter 29 assumes perfectly inflexible prices (and wages) and thus is a model in which prices are not just sticky but completely stuck. By contrast, the aggregate demand–aggregate supply model presented in Chapter 30 allows for flexible prices (with or without flexible wages) and is therefore useful for understanding how the economy behaves over longer periods of time.

As you study these various models, keep in mind that we need different models precisely because the economy behaves so differently depending on how much time has passed after a demand shock. The differences in behavior result from the fact that prices go from stuck in the extreme short run to fully flexible in the long run. Using different models for different stages in this process gives us much better insights into not only how economies actually behave but also how various government and central bank policies may have different effects in the short run when prices are fixed versus the long run when prices are flexible.

Where will we go from here? In the remainder of Part 7, we examine how economists measure GDP and why GDP has expanded over time. Then, we discuss the terminology of business cycles and explore the measurement and types of unemployment and inflation. At that point you will be well-prepared to examine the economic models, monetary considerations, and stabilization policies that lie at the heart of macroeconomics.

QUICK REVIEW 24.5

- Many commodity prices are extremely flexible and change constantly, but other prices in the economy change only very infrequently.
- Some prices are inflexible in order to please retail customers, others because rival firms are afraid that price changes may trigger a price war.
- Prices tend to become more flexible over time, so that as time passes, the economy can react to demand shocks with price changes as well as with output and employment changes.
- Different macroeconomics models are required for the short run, during which prices are inflexible (so that demand shocks lead almost exclusively to output and employment changes), and for longer periods, during which prices become increasingly flexible (so that demand shocks lead more to price changes rather than output and employment changes).

SUMMARY

LO24.1 Explain why economists focus on GDP, inflation, and unemployment when assessing the health of an entire economy.

Macroeconomics studies long-run economic growth and short-run economic fluctuations.

Macroeconomists focus their attention on three key economic statistics: real GDP, unemployment, and inflation. Real GDP measures the value of all final goods and services produced in a country during a specific period of time. The unemployment rate measures the percentage of all workers who are not able to find paid employment despite being willing and able to work at currently available wages. The inflation rate measures the extent to which the overall level of prices is rising in the economy.

LO24.2 Discuss why sustained increases in living standards are a historically recent phenomenon.

Before the Industrial Revolution, living standards did not show any sustained increases over time. Economies grew, but any increase in output tended to be offset by an equally large increase in the population, so that the amount of output per person did not rise. By contrast, since the Industrial Revolution began in the late 1700s, many nations have experienced modern economic growth in which output grows faster than population—so that standards of living rise over time.

LO24.3 Identify why saving and investment are key factors in promoting rising living standards.

Macroeconomists believe that one of the keys to modern economic growth is the promotion of saving and investment (for economists, the purchase of capital goods). Investment activities increase the economy's future potential output level. But investment must be funded by saving, which is only possible if people are willing to reduce current consumption. Consequently, individuals and society face a trade-off between current consumption and future consumption since the only way to fund the investment necessary to increase future consumption is by reducing current consumption in order to gather the savings necessary to fund that investment. Banks and other financial institutions help to convert saving into investment by taking the savings generated by households and lending it to businesses that wish to make investments.

LO24.4 Describe why economists believe that "shocks" and "sticky prices" are responsible for short-run fluctuations in output and employment.

Expectations have an important effect on the economy for two reasons. First, if people and businesses are more positive about the future, they will save and invest more. Second, individuals and firms must make adjustments to shocks—situations in which expectations are unmet and the future does not turn out the way people were expecting. In particular, shocks often imply situations where the quantity supplied of a given good or service does not equal the quantity demanded of that good or service.

If prices were always flexible and capable of rapid adjustment, then dealing with situations in which quantities demanded did not equal quantities supplied would always be easy since prices could simply adjust to the market equilibrium price at which quantities demanded equal quantities supplied. Unfortunately, real-world prices are often inflexible (or "sticky") in the short run so that the only way for the economy to adjust to such situations is through changes in output levels.

Sticky prices combine with shocks to drive short-run fluctuations in output and employment. Consider a negative demand shock in which demand is unexpectedly low. Because prices are fixed, the lower-than-expected demand will result in unexpectedly slow sales. This will cause inventories to increase. If demand remains low for an extended period of time, inventory levels will become too high and firms will have to cut output and lay off workers. Thus, when prices are inflexible, the economy adjusts to unexpectedly low demand through changes in output and employment rather than through changes in prices (which are not possible when prices are inflexible).

LO24.5 Characterize the degree to which various prices in the economy are sticky.

Prices are inflexible in the short run for various reasons, two of which are discussed in this chapter. First, firms often attempt to set and maintain stable prices to please customers who like predictable prices because they make for easy planning (and who might become upset if prices were volatile). Second, a firm with just a few competitors may be reluctant to cut its price due to the fear of starting a price war, a situation in which its competitors retaliate by cutting their prices as well—thereby leaving the firm worse off than it was to begin with.

LO24.6 Explain why the greater flexibility of prices as time passes causes economists to utilize different macroeconomic models for different time horizons.

Price stickiness moderates over time. As a result, economists have found it sensible to build separate economic models for different time horizons. For instance, some models are designed to reflect the high degree of price inflexibility that occurs in the immediate short run, while other models reflect the high degree of price flexibility that occurs in the long run. The different models allow economists to have a better sense for how various government policies will affect the economy in the short run when prices are inflexible versus the long run when prices are flexible.

TERMS AND CONCEPTS

business cycle
recession
real GDP (gross domestic product)
nominal GDP
unemployment
inflation
modern economic growth
saving
investment
financial investment
economic investment
expectations
shocks
demand shocks
supply shocks
inventory
inflexible prices ("sticky prices")
flexible prices

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Why do you think macroeconomists focus on just a few key statistics when trying to understand the health and trajectory of an economy? Would it be better to try to examine all possible data? **LO24.1**
2. Consider a nation in which the volume of goods and services is growing by 5 percent per year. What is the likely impact of this high rate of growth on the power and influence of its government relative to other countries experiencing slower rates of growth? What about the effect of this 5 percent growth on the nation's living standards? Will these also necessarily grow by 5 percent per year, given population growth? Why or why not? **LO24.2**
3. Did economic output start growing faster than population from the beginning of the human inhabitation of the earth? When did modern economic growth begin? Have all of the world's nations experienced the same extent of modern economic growth? **LO24.2**
4. Why is there a trade-off between the amount of consumption that people can enjoy today and the amount of consumption that they can enjoy in the future? Why can't people enjoy more of both? How does saving relate to investment and thus to economic growth? What role do banks and other financial institutions play in aiding the growth process? **LO24.3**
5. How does investment as defined by economists differ from investment as defined by the general public? What would happen to the amount of economic investment made today if firms expected the future returns to such investment to be very low? What if firms expected future returns to be very high? **LO24.3**
6. Why, in general, do shocks force people to make changes? Give at least two examples from your own experience. **LO24.4**
7. Catalog companies are committed to selling at the prices printed in their catalogs. If a catalog company finds its inventory of sweaters rising, what does that tell you about the demand for sweaters? Was it unexpectedly high, unexpectedly low, or as expected? If the company could change the price of sweaters, would it raise the price, lower the price, or keep the price the same? Given that the company cannot change the price of sweaters, consider the number of sweaters it orders each month from the company that makes its sweaters. If inventories become very high, will the catalog company increase, decrease, or keep orders the same? Given what the catalog company does with its orders, what is likely to happen to employment and output at the sweater manufacturer? **LO24.4**
8. Are all prices in the economy equally inflexible? Which ones show large amounts of short-run flexibility? Which ones show a great deal of inflexibility even over months and years? **LO24.5**
9. Why do many firms strive to maintain stable prices? **LO24.5**
10. Do prices tend to become more or less flexible as time passes? If there is a trend, how does it affect macroeconomists' choice of models? **LO24.6**
11. **LAST WORD** How do the Minsky and Austrian explanations for the causes of the Great Recession differ? Explain how the proponents of government stimulus believe that it will affect aggregate demand and employment (be specific!). How might government stimulus possibly slow rather than accelerate a recovery?

REVIEW QUESTIONS

1. An increase in ______ GDP guarantees that more goods and services are being produced by an economy. **LO24.1**
   a. Nominal.
   b. Real.
2. True or False. The term *economic investment* includes purchasing stocks, bonds, and real estate. **LO24.3**
3. If an economy has sticky prices and demand unexpectedly increases, you would expect the economy's real GDP to: **LO24.4**
   a. Increase.
   b. Decrease.
   c. Remain the same.

4. If an economy has fully flexible prices and demand unexpectedly increases, you would expect that the economy's real GDP would tend to: **LO24.4**
   a. Increase.
   b. Decrease.
   c. Remain the same.
5. If the demand for a firm's output unexpectedly decreases, you would expect that its inventory would: **LO24.4**
   a. Increase.
   b. Decrease.
   c. Remain the same.
   d. Increase or remain the same, depending on whether prices are sticky.
6. True or False. Because price stickiness only matters in the short run, economists are comfortable using just one macroeconomic model for all situations. **LO24.6**

PROBLEMS

1. Suppose that the annual rates of growth of real GDP of Econoland over a five-year period were sequentially as follows: 3 percent, 1 percent, –2 percent, 4 percent, and 5 percent. What was the average of these growth rates in Econoland over these 5 years? What term would economists use to describe what happened in year 3? If the growth rate in year 3 had been a positive 2 percent rather than a negative 2 percent, what would have been the average growth rate? **LO24.1**
2. Suppose that Glitter Gulch, a gold mining firm, increased its sales revenues on newly mined gold from \$100 million to \$200 million between one year and the next. Assuming that the price of gold increased by 100 percent over the same period, by what numerical amount did Glitter Gulch's real output change? If the price of gold had not changed, what would have been the change in Glitter Gulch's real output? **LO24.1**
3. A mathematical approximation called the rule of 70 tells us that the number of years that it will take something that is growing to double in size is approximately equal to the number 70 divided by its percentage rate of growth. Thus, if Mexico's real GDP per person is growing at 7 percent per year, it will take about 10 years (= 70/7) to double. Apply the rule of 70 to solve the following problem. Real GDP per person in Mexico in 2005 was about \$11,000 per person, while it was about \$44,000 per person in the United States. If real GDP per person in Mexico grows at the rate of 5 percent per year, about how long will it take Mexico's real GDP per person to reach the level that the United States was at in 2005? (Hint: How many times would Mexico's 2005 real GDP per person have to double to reach the United States' 2005 real GDP per person?) **LO24.2**
4. Assume that a national restaurant firm called BBQ builds 10 new restaurants at a cost of \$1 million per restaurant. It outfits each restaurant with an additional \$200,000 of equipment and furnishings. To help partially defray the cost of this expansion, BBQ issues and sells 200,000 shares of stock at \$30 per share. What is the amount of economic investment that has resulted from BBQ's actions? How much purely financial investment took place? **LO24.3**
5. Refer to Figure 24.1b and assume that price is fixed at \$37,000 and that Buzzer Auto needs 5 workers for every 1 automobile produced. If demand is D_M and Buzzer wants to perfectly match its output and sales, how many cars will Buzzer produce and how many workers will it hire? If instead, demand unexpectedly falls from D_M to D_L, how many fewer cars will Buzzer sell? How many fewer workers will it need if it decides to match production to these lower sales? **LO24.4**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **25**

Measuring Domestic Output and National Income

Learning Objectives:

LO25.1 Explain how gross domestic product (GDP) is defined and measured.

LO25.2 Describe how expenditures on goods and services can be summed to determine GDP.

LO25.3 Explain how GDP can be determined by summing up all of the incomes that were derived from producing the economy's output of goods and services.

LO25.4 Describe the relationships among GDP, net domestic product, national income, personal income, and disposable income.

LO25.5 Discuss the nature and function of a GDP price index, and describe the difference between nominal GDP and real GDP.

LO25.6 List and explain some limitations of the GDP measure.

"Disposable Income Flat." "Personal Consumption Surges." "Investment Spending Stagnates." "GDP Up 4 Percent." These headlines, typical of those found on Yahoo! Finance or in *The Wall Street Journal*, give knowledgeable readers valuable information on the state of the economy. This chapter will help you interpret such headlines and understand the stories reported under them. Specifically, it will help you become familiar with the vocabulary and methods of national income accounting. Such accounting enables economists to measure the long-run rate of economic growth and identify the recessions and expansions associated with the economic ups and downs known as the business cycle. In addition, the terms and ideas that you encounter in this chapter will provide a needed foundation for the macroeconomic models found in subsequent chapters.

Assessing the Economy's Performance

LO25.1 Explain how gross domestic product (GDP) is defined and measured.

National income accounting measures the economy's overall performance. It does for the economy as a whole what private accounting does for the individual firm or for the individual household.

A business firm measures its flows of income and expenditures regularly—usually every 3 months or once a year. With that information in hand, the firm can gauge its economic health. If things are going well and profits are good, the accounting data can be used to explain that success. Were costs down? Was output up? Have market prices risen? If things are going badly and profits are poor, the firm may be able to identify the reason by studying the record over several accounting periods. All this information helps the firm's managers plot their future strategy.

National income accounting operates in much the same way for the economy as a whole. The Bureau of Economic Analysis (BEA), an agency of the Commerce Department, compiles the National Income and Product Accounts (NIPA) for the U.S. economy. This accounting enables economists and policymakers to:

- Assess the health of the economy by comparing levels of production at regular intervals.
- Track the long-run course of the economy to see whether it has grown, been constant, or declined.
- Formulate policies that will safeguard and improve the economy's health.

Gross Domestic Product

The primary measure of the economy's performance is its annual total output of goods and services or, as it is called, its *aggregate output*. There are several ways to measure aggregate output depending upon how one wishes to define "an economy." For instance, should the value of the cars produced at a Toyota plant in Ohio count as part of the output of the U.S. economy because they are made within the United States or as part of the Japanese economy because Toyota is a Japanese company? As mentioned in Chapter 24, **gross domestic product (GDP)** defines aggregate output as the dollar value of all final goods and services produced within the borders of a country during a specific period of time, typically a year. Under this definition, the value of the cars produced at the Toyota factory in Ohio clearly count as part of U.S. aggregate output rather than Japanese aggregate output because the cars are made within the borders of the United States.[1]

TABLE 25.1 Comparing Heterogeneous Output by Using Money Prices

Year	Annual Output	Market Value
1	3 sofas and 2 computers	3 at \$500 + 2 at \$2,000 = \$5,500
2	2 sofas and 3 computers	2 at \$500 + 3 at \$2,000 = \$7,000

A Monetary Measure

By necessity, GDP is a *monetary measure*. To see why, suppose that the economy produces three sofas and two computers in year 1 and two sofas and three computers in year 2. In which year is output greater? We can't answer that question until we attach a price tag to each of the two products to indicate how society evaluates their relative worth.

That's what GDP does. It measures the value of output in monetary terms. Without such a measure we would have no way of comparing the relative values of the vast number of goods and services produced in different years. In Table 25.1 the price of sofas is \$500 and the price of computers is \$2,000. GDP would gauge the output of year 2 (\$7,000) as greater than the output of year 1 (\$5,500) because society places a higher monetary value on the output of year 2. Society is willing to pay \$1,500 more for the combination of goods produced in year 2 than for the combination of goods produced in year 1.

[1]In contrast to GDP, U.S. gross *national* product (GNP) consists of the total value of all the final goods and services produced by American-supplied resources, whether those goods and services are produced within the borders of the United States or abroad. The U.S. switched from GNP to GDP accounting in 1992 to match the type of accounting used by other countries worldwide.

Avoiding Multiple Counting

To measure aggregate output accurately, all goods and services produced in a particular year must be counted once and only once. Because most products go through a series of production stages before they reach the market, some of their components are bought and sold many times. To avoid counting those components each time, GDP includes only the market value of *final goods* and ignores *intermediate goods* altogether.

Intermediate goods are products that are purchased for resale or further processing or manufacturing. **Final goods** are products that are purchased by their end users. Crude oil is an intermediate good; gasoline used for personal transportation is a final good. Steel beams are intermediate goods; completed high-rise apartments are final goods. Lettuce, carrots, and vinegar in restaurant salads are intermediate goods; restaurant salads are final goods. Other examples of final goods are sunglasses bought by consumers, assembly machinery purchased by businesses, surveillance satellites bought by government, and smart phones purchased by foreign buyers.

Why is the value of final goods included in GDP but the value of intermediate goods excluded? Because the value of final goods already includes the value of all the intermediate goods that were used in producing them. Including the value of intermediate goods would amount to **multiple counting,** and that would distort the value of GDP.

To see why, suppose that five stages are needed to manufacture a wool coat and get it to the consumer—the final user. Table 25.2 shows that firm A, a sheep ranch, sells $120 worth of wool to firm B, a wool processor. Firm A pays out the $120 in wages, rent, interest, and profit. Firm B processes the wool and sells it to firm C, a coat manufacturer, for $180. What does firm B do with the $180 it receives? It pays $120 to firm A for the wool and uses the remaining $60 to pay wages, rent, interest, and profit for the resources used in processing the wool. Firm C, the manufacturer, sells the coat to firm D, a wholesaler, which sells it to firm E, a retailer. Then at last a consumer, the final user, comes in and buys the coat for $350.

How much of these amounts should we include in GDP to account for the production of the coat? Just $350, the value of the final product. The $350 includes all the intermediate transactions leading up to the product's final sale. Including the sum of all the intermediate sales, $1,140, in GDP would amount to multiple counting. The production and sale of the final coat generated just $350 of output, not $1,140.

Alternatively, we could avoid multiple counting by measuring and cumulating only the *value added* at each stage. **Value added** is the market value of a firm's output *less* the value of the inputs the firm has bought from others. At each stage, the difference between what a firm pays for inputs and what it receives from selling the product made from those inputs is paid out as wages, rent, interest, and profit. Column 3 of Table 25.2 shows that the value added by firm B is $60, the difference between the $180 value of its output and the $120 it paid for the input from firm A. We find the total value of the coat by adding together all the values added by the five firms. Similarly, by calculating and summing the values added to all the goods and services produced by all firms in the economy, we can find the market value of the economy's total output—its GDP.

GDP Excludes Nonproduction Transactions

Although many monetary transactions in the economy involve final goods and services, many others do not. These nonproduction transactions must be excluded from GDP because they have nothing to do with the generation of final goods. *Nonproduction transactions* are

TABLE 25.2 Value Added in a Five-Stage Production Process

(1) Stage of Production	(2) Sales Value of Materials or Product	(3) Value Added
	$ 0	
Firm A, sheep ranch	120	$120 (= $120 − $ 0)
Firm B, wool processor	180	60 (= 180 − 120)
Firm C, coat manufacturer	220	40 (= 220 − 180)
Firm D, clothing wholesaler	270	50 (= 270 − 220)
Firm E, retail clothier	350	80 (= 350 − 270)
Total sales values	$1,140	
Value added (total income)		$350

of two types: purely financial transactions and secondhand sales.

Financial Transactions Purely financial transactions include the following:

- ***Public transfer payments*** These are the social security payments, welfare payments, and veterans' payments that the government makes directly to households. Since the recipients contribute nothing to *current production* in return, to include such payments in GDP would be to overstate the year's output.
- ***Private transfer payments*** Such payments include, for example, the money that parents give children or the cash gifts given during the holidays. They produce no output. They simply transfer funds from one private individual to another and consequently do not enter into GDP.
- ***Stock market transactions*** The buying and selling of stocks (and bonds) is just a matter of swapping bits of paper. Stock market transactions create nothing in the way of current production and are not included in GDP. Payments for the services provided by a stockbroker *are* included, however, because their services are currently provided and are thus a part of the economy's current output of goods and services.

Secondhand Sales Secondhand sales contribute nothing to current production and for that reason are excluded from GDP. Suppose you sell your 2005 Ford Mustang to a friend; that transaction would be ignored in reckoning this year's GDP because it generates no current production. The same would be true if you sold a brand-new Mustang to a neighbor a week after you purchased it.

Two Ways of Looking at GDP: Spending and Income

Let's look again at how the market value of total output—or of any single unit of total output—is measured. Given the data listed in Table 25.2, how can we measure the market value of a coat?

One way is to see how much the final user paid for it. That will tell us the market value of the final product. Or we can add up the entire wage, rental, interest, and profit incomes that were created in producing the coat. The second approach is the value-added technique used in Table 25.2.

The final-product approach and the value-added approach are two ways of looking at the same thing. What is spent on making a product is income to those who helped make it. If $350 is spent on manufacturing a coat, then $350 is the total income derived from its production.

We can look at GDP in the same two ways. We can view GDP as the sum of all the money spent in buying it. That is the *output approach*, or **expenditures approach.** Or we can view GDP in terms of the income derived or created from producing it. That is the *earnings* or *allocations approach*, or the **income approach.**

As illustrated in Figure 25.1, we can determine GDP for a particular year either by adding up all that was spent to buy total output or by adding up all the money that was derived as income from its production. Buying (spending money) and selling (receiving income) are two aspects of the same transaction. On the expenditures side of GDP, all final goods produced by the economy are bought either by three domestic sectors (households, businesses, and government) or by foreign buyers. On the income side (once certain statistical adjustments are made), the

FIGURE 25.1 The expenditures and income approaches to GDP. There are two general approaches to measuring gross domestic product. We can determine GDP as the value of output by summing all expenditures on that output. Alternatively, with some modifications, we can determine GDP by adding up all the components of income arising from the production of that output.

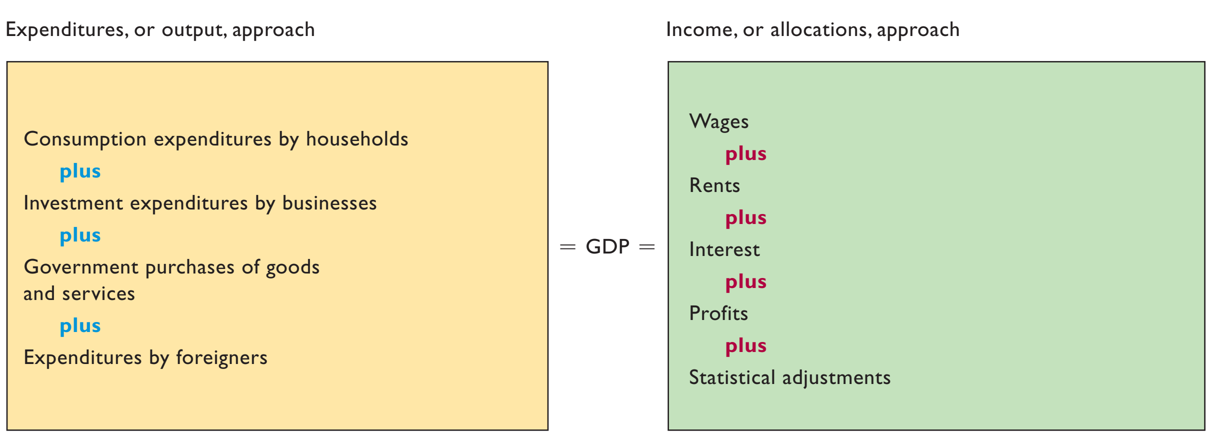


TABLE 25.3 **Accounting Statement for the U.S. Economy, 2012 (in Billions)***

Receipts: Expenditures Approach		Allocations: Income Approach	
Sum of:		Sum of:	
Personal consumption expenditures (C)	$11,150	Compensation of employees	$8,612
Gross private domestic investment (I_g)	2,475	Rents	541
Government purchases (G)	3,167	Interest	440
Net exports (X_n)	−547	Proprietors' income	1,225
		Corporate profits	2,031
		Taxes on production and imports	1,123
		Equals:	
		National income	**$13,972**
		National income	$13,972
		Less: Net foreign factor income	253
		Plus: Consumption of fixed capital	2,543
		Plus: Statistical discrepancy	−17
Equals:		*Equals:*	
Gross domestic product	**$16,245**	***Gross domestic product***	**$16,245**

*Some of the items in the Allocations column combine related categories that appear in the more detailed accounts. All data are subject to government revision.

Source: Bureau of Economic Analysis, **www.bea.gov**.

total receipts acquired from the sale of that total output are allocated to the suppliers of resources as wage, rent, interest, and profit.

Table 25.3 shows U.S. GDP for the year 2012 totaled up using both the expenditures approach (on the left side) and the income approach (on the right side). As you would expect, both methods reach the same conclusion: U.S. GDP in 2012 was $16,245 billion.

We will now go through both approaches in detail. Doing so will help you better understand both methods and, in particular, why the income side of Table 25.3 looks substantially more complicated than the income side of Figure 25.1.

The Expenditures Approach

LO25.2 Describe how expenditures on goods and services can be summed to determine GDP.

To determine GDP using the expenditures approach, we add up all the spending on final goods and services that has taken place throughout the year. National-income accountants use precise terms for the types of spending listed on the left side of Figure 25.1.

Personal Consumption Expenditures (C)

What we have called "consumption expenditures by households," the national income accountants call **personal consumption expenditures.** This term covers all expenditures by households on goods and services.

In a typical year, roughly 10 percent of these personal consumption expenditures are on **durable goods**—products that have expected lives of three years or more. Such goods include new automobiles, furniture, and refrigerators. Another 30 percent are on **nondurable goods**—products with less than three years of expected life. Included are goods like food, clothing, and gasoline. About 60 percent of personal consumption expenditures are on **services**—the work done by lawyers, hair stylists, doctors, mechanics, and other service providers. Because of this high percentage, economists sometimes refer to the U.S. economy as a *service economy*. National income accountants combine the household spending on durable goods, nondurable goods, and services and use the symbol C to designate the personal consumption expenditures component of GDP.

Gross Private Domestic Investment (I_g)

Under the heading **gross private domestic investment,** the accountants include the following items:

- All final purchases of machinery, equipment, and tools by business enterprises.
- All construction.
- Changes in inventories.
- Money spent on research and development (R&D) or for the creation of new works of art, music, writing, film, and so on.

Notice that this list, except for the first item, includes more than we have meant by "investment" so far. The second item includes residential construction as well as the construction of new factories, warehouses, and stores. Why do the accountants regard residential construction as investment rather than consumption? Because apartment buildings and houses, like factories and stores, earn income when they are rented or leased. Owner-occupied houses are treated as investment goods because they *could be* rented to bring in an income return. So the national income accountants treat all residential construction as investment.

Increases in inventories (unsold goods) are considered to be investment because they represent, in effect, "unconsumed output." For economists, all new output that is not consumed is, by definition, capital. An increase in inventories is an addition (although perhaps temporary) to the stock of capital goods, and such additions are precisely how we define investment.

Starting in 2013, the NIPA accountants who compile U.S. GDP statistics began to include expenditures on R&D as well as money spent to develop new works of writing, art, music, and software as a form of investment. They did so because a country's stock of "capital goods" useful in producing output can be thought of as including not only tangible pieces of physical capital like fiber optic networks and factories but also useful ideas that increase the economy's ability to produce goods and services.

Software is a great example, as it is merely sets of instructions for telling computers what to do. But without those instructions, computers would be useless. So spending on software as well as on R&D and other intellectual activities that improve the economy's stock of "know-how" are now counted as investment.

To make it possible to compare GDP numbers across time, the accountants have gone back and applied the new, more comprehensive definition of investment all the way back to 1929. The numbers for U.S. GDP for the year 2012 that are used in this chapter incorporate the revised definition of investment.

Positive and Negative Changes in Inventories We need to look at changes in inventories more closely. Inventories can either increase or decrease over some period. Suppose they increased by $10 billion between December 31, 2012, and December 31, 2013. Therefore, in 2013 the economy produced $10 billion more output than people purchased. We need to count all output produced in 2013 as part of that year's GDP, even though some of it remained unsold at the end of the year. This is accomplished by including the $10 billion increase in inventories as investment in 2013. That way the expenditures in 2013 will correctly measure the output produced that year.

Alternatively, suppose that inventories decreased by $10 billion in 2013. This "drawing down of inventories" means that the economy sold $10 billion more of output in 2013 than it produced that year. It did this by selling goods produced in prior years—goods already counted as GDP in those years. Unless corrected, expenditures in 2013 will overstate GDP for 2013. So in 2013 we consider the $10 billion decline in inventories as "negative investment" and subtract it from total investment that year. Thus, expenditures in 2013 will correctly measure the output produced in 2013.

Noninvestment Transactions So much for what investment *is*. You also need to know what it *isn't*. For economists and NIPA accountants, investment does *not* include noninvestment transactions such as the transfer of paper assets (stocks, bonds) or the resale of tangible assets (houses, jewelry, boats). Such financial transactions merely transfer the ownership of existing assets. The investment in the GDP accounts is economic investment—the creation of *new* capital assets. The mere transfer (sale) of claims to existing capital goods does not produce new capital goods. Therefore such transactions (so-called financial investments) are not included as investment in the GDP accounts.

Gross Investment versus Net Investment As we have seen, the category gross private domestic investment includes (1) all final purchases of machinery, equipment, and tools; (2) all construction; (3) changes in inventories; and (4) spending on R&D and other activities that expand the economy's stock of technology and know-how. The words "private" and "domestic" mean that we are speaking of spending by private businesses, not by government (public) agencies, and that the investment is taking place inside the country, not abroad.

The word "gross" means that we are referring to *all* investment goods—both those that replace machinery, equipment, and buildings that were used up (worn out or made obsolete) in producing the current year's output and any net additions to the economy's stock of capital. Gross investment includes investment in replacement capital *and* in added capital.

In contrast, **net private domestic investment** includes *only* investment in the form of added capital. The amount of capital that is used up over the course of a year is called *depreciation*. So

Net investment = gross investment − depreciation

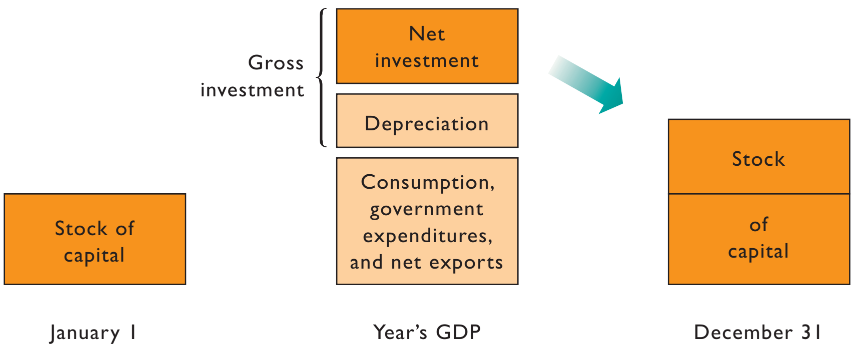


FIGURE 25.2 Gross investment, depreciation, net investment, and the stock of capital. When gross investment exceeds depreciation during a year, net investment occurs. This net investment expands the stock of private capital from the beginning of the year to the end of the year by the amount of the net investment. Other things equal, the economy's production capacity expands.

In typical years, gross investment exceeds depreciation. Thus net investment is positive and the nation's stock of capital rises by the amount of net investment. As illustrated in Figure 25.2, the stock of capital at the end of the year exceeds the stock of capital at the beginning of the year by the amount of net investment.

Gross investment need not always exceed depreciation, however. When gross investment and depreciation *are equal*, net investment is zero and there is no change in the size of the capital stock. When gross investment *is less than* depreciation, net investment is negative. The economy then is *disinvesting*—using up more capital than it is producing—and the nation's stock of capital shrinks. That happened in the Great Depression of the 1930s.

National income accountants use the symbol I for private domestic investment spending. To differentiate between gross investment and net investment, they add either the subscript g or the subscript n. But it is gross investment, I_g, that they use when tallying up GDP.

Government Purchases (G)

The third category of expenditures in the national income accounts is **government purchases,** officially labeled "government consumption expenditures and gross investment." These expenditures have three components: (1) expenditures for goods and services that government consumes in providing public services; (2) expenditures for *publicly owned capital* such as schools and highways, which have long lifetimes; and (3) government expenditures on R&D and other activities that increase the economy's stock of know-how. Government purchases (federal, state, and local) include all government expenditures on final goods and all direct purchases of resources, including labor. It does *not* include government transfer payments because, as we have seen, they merely transfer government receipts to certain households and generate no production of any sort. National income accountants use the symbol G to signify government purchases.

Net Exports (X_n)

International trade transactions are a significant item in national income accounting. But when calculating U.S. GDP, we must keep in mind that we want to total up only those expenditures that are used to purchase goods and services produced *within the borders of the United States*. Thus, we must add in the value of exports, X, since exports are by definition goods and services produced within the borders of the United States. Don't be confused by the fact that the expenditures made to buy our exports are made by foreigners. The definition of GDP does not care about *who* is making expenditures on U.S.-made goods and services—only that the goods and services that they buy are made within the borders of the United States. Thus, foreign spending on our exports *must* be included in GDP.

At this point, you might incorrectly think that GDP should be equal to the sum of $C + I_g + G + X$. But this sum overstates GDP. The problem is that, once again, we must consider only expenditures made on *domestically produced* goods and services. As it stands, C, I_g, and G count expenditures on consumption, investment, and government purchases *regardless* of where those goods and services are made. Crucially, not all of the C, I_g, or G expenditures are for domestically produced goods and services. Some of the expenditures are for imports—goods and services produced outside of the United States. Because we wish to count *only* the part of C, I_g, and G that goes to purchasing domestically produced goods and services, we must subtract the spending that goes to imports, M. That subtraction yields the correct formula for calculating gross domestic product: GDP = $C + I_g + G + X - M$.

Accountants simplify this formula for GDP by defining **net exports,** X_n, to be equal to exports minus imports:

$$\text{Net exports } (X_n) = \text{exports } (X) - \text{imports } (M)$$

CONSIDER THIS . . .

Stocks versus Flows

An analogy of a reservoir is helpful in thinking about a nation's capital stock, investment, and depreciation. Picture a reservoir that has water flowing in from a river and flowing out from an outlet after it passes through turbines. The volume of water in the reservoir *at any particular point in time* is a "stock." In contrast, the inflow from the river and outflow from the outlet are "flows."

The volume or stock of water in the reservoir will rise if the weekly inflow exceeds the weekly outflow. It will fall if the inflow is less than the outflow. And it will remain constant if the two flows are equal.

Now let's apply this analogy to the stock of capital, gross investment, and depreciation. The stock of capital is the total capital in place at any point in time and is analogous to the level of water in the reservoir. Changes in this capital stock over some period, for example, one year, depend on *gross investment* and *depreciation*. Gross investment (analogous to the reservoir inflow) is an addition of capital goods and therefore adds to the stock of capital, while depreciation (analogous to the reservoir outflow) is the using up of capital and thus subtracts from the capital stock. The capital stock increases when gross investment exceeds depreciation, declines when gross investment is less than depreciation, and remains the same when gross investment and depreciation are equal.

Alternatively, the stock of capital increases when *net investment* (gross investment *minus* depreciation) is positive. When net investment is negative, the stock of capital declines, and when net investment is zero, the stock of capital remains constant.

Using this definition of net exports, the formula for gross domestic product simplifies to,

$$\text{GDP} = C + I_g + G + X_n$$

The left side of Table 25.3 shows that in 2012 Americans spent \$560 billion more on imports than foreigners spent on U.S. exports. That is, net exports in 2012 were a *minus* \$560 billion.

Putting It All Together: GDP = $C + I_g + G + X_n$

Taken together, the four categories of expenditures provide a measure of the market value of a specific year's total output—its GDP. For the United States in 2012, the left side of Table 25.3 indicates that

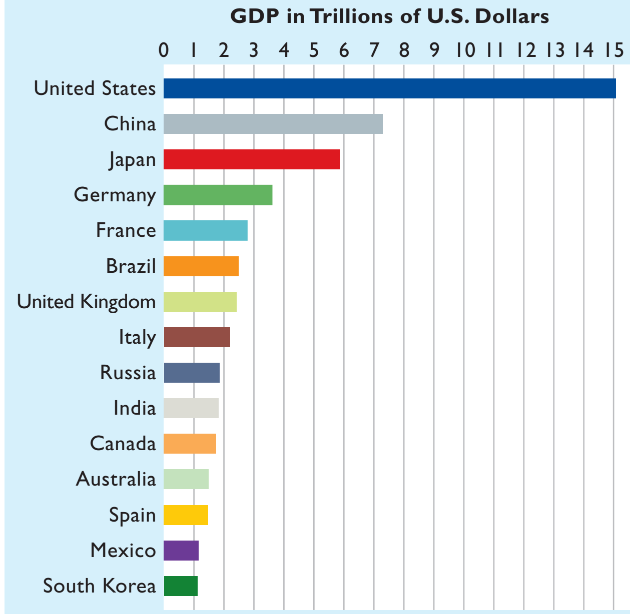
$$\text{GDP} = \$11{,}150 + \$2{,}475 + \$3{,}167 - 547 \text{ billion}$$

Global Perspective 25.1 lists the GDPs of several countries. The values of GDP are converted to dollars using international exchange rates.

GLOBAL PERSPECTIVE 25.1

Comparative GDPs in Trillions of U.S. Dollars, Selected Nations, 2011

The United States, China, and Japan have the world's highest GDPs. The GDP data charted below have been converted to U.S. dollars via international exchange rates.



Source: International Monetary Fund, **www.imf.org**.

The Income Approach

LO25.3 Explain how GDP can be determined by summing up all of the incomes that were derived from producing the economy's output of goods and services.

The right side of Table 25.3 shows how 2012's expenditures of \$16,245 billion were allocated as income to those responsible for producing the output. It would be simple if we could say that the entire amount of expenditures flowed back to them in the form of wages, rent, interest, and profit. But some expenditures flow to other recipients (such as the government) or to other uses (such as paying to replace the capital goods that have worn out while producing this year's

GDP). These must be accounted for to balance the expenditures and income sides of the overall account. We will begin by looking at the items that make up *national income*.

Compensation of Employees

By far the largest share of national income—$8,612 billion in 2012—was paid as wages and salaries by business and government to their employees. That figure also includes wage and salary supplements, in particular, payments by employers into social insurance and into a variety of private pension, health, and welfare funds for workers.

Rents

Rents consist of the income received by the households and businesses that supply property resources. They include the monthly payments tenants make to landlords and the lease payments corporations pay for the use of office space. The figure used in the national accounts is *net* rent—gross rental income minus depreciation of the rental property.

Interest

Interest consists of the money paid by private businesses to the suppliers of loans used to purchase capital. It also includes such items as the interest households receive on savings deposits, certificates of deposit (CDs), and corporate bonds.

Proprietors' Income

What we have loosely termed "profits" is broken down by the national income accountants into two accounts: proprietors' income, which consists of the net income of sole proprietorships, partnerships, and other unincorporated businesses; and corporate profits. Proprietors' income flows to the proprietors.

Corporate Profits

Corporate profits are the earnings of corporations. National income accountants subdivide corporate profits into three categories:

- ***Corporate income taxes*** These taxes are levied on corporations' profits. They flow to the government.
- ***Dividends*** These are the part of after-tax profits that corporations choose to pay out, or distribute, to their stockholders. They thus flow to households—the ultimate owners of all corporations.
- ***Undistributed corporate profits*** Any after-tax profits that are not distributed to shareholders are saved, or retained, by corporations to be invested later in new plants and equipment. Undistributed corporate profits are also called *retained earnings*.

Taxes on Production and Imports

The account called **taxes on production and imports** includes general sales taxes, excise taxes, business property taxes, license fees, and customs duties. Why do national income accountants add these indirect business taxes to wages, rent, interest, and profits in determining national income? The answer is, "to account for expenditures that are diverted to the government." Consider an item that would otherwise sell for $1 but costs $1.05 because the government has imposed a 5 percent sales tax. When this item is purchased, consumers will expend $1.05 to buy it. But only $1 will go to the seller (who will then distribute it as income in the form of wages, rent, interest, and profit in order to compensate resource providers). The remaining 5 cents will flow as revenue to the government. The GDP accountants handle the extra 5 cents by placing it into the category called "Taxes on Production and Imports" and loosely consider it to be "income" to government.

From National Income to GDP

We have just shown that expenditures on final goods and services flow either as income to private citizens or as "income" to government. As a result, **national income** is the total of all sources of private income (employee compensation, rents, interest, proprietors' income, and corporate profits) plus government revenue from taxes on production and imports. National income is all the income that flows to American-supplied resources, whether here or abroad, plus taxes on production and imports. But notice that the figure for national income shown in Table 25.3—$13,972 billion—is less than GDP as reckoned by the expenditures approach shown on the left side of the table. The two sides of the accounting statement are brought into balance by subtracting one item from national income and adding two others.

Net Foreign Factor Income First, we need to make a slight adjustment in "national" income versus "domestic" income. National income includes the total income of Americans, whether it was earned in the United States or abroad. But GDP is a measure of *domestic* output—total output produced within the United States regardless of the nationality of those who provide the resources. So in moving from national income to GDP, we must take out the income Americans gain from supplying resources abroad and add in the income that foreigners gain by supplying resources in the United States. That process provides *net foreign factor income*. In 2012, net foreign factor income was \$253 billion, meaning that American-owned resources earned \$253 billion more in other countries than foreign-owned resources earned in the United States. Because this \$253 billion is earnings of Americans, it is included in U.S. national income. But this income is not part of U.S. domestic income because it reflects earnings from output produced in other nations. It is part of those nations' domestic income, derived from production of their domestic output. Thus, we subtract net foreign factor income from U.S. national income to stay on the correct path to use the income approach to determine the value of *U.S. domestic output* (output produced within U.S. borders).

Consumption of Fixed Capital Next, we must recognize that the useful lives of private capital equipment (such as bakery ovens or automobile assembly lines) extend far beyond the year in which they were produced. To avoid understating profit and income in the year of purchase and to avoid overstating profit and income in succeeding years, the cost of such capital must be allocated over its lifetime. The amount allocated is an estimate of how much of the capital is being used up each year. It is called *depreciation*. Accounting for depreciation results in a more accurate statement of profit and income for the economy each year. Publicly owned capital, such as courthouses and bridges, also requires a depreciation allowance in the national income accounts.

The huge depreciation charge made against private and publicly owned capital each year is called **consumption of fixed capital** because it is the allowance for capital that has been "consumed" in producing the year's GDP. It is the portion of GDP that is set aside to pay for the ultimate replacement of those capital goods.

The money allocated to consumption of fixed capital (the depreciation allowance) is a cost of production and thus included in the gross value of output. But this money is not available for other purposes, and, unlike other costs of production, it does not add to anyone's income. So it is not included in national income. We must therefore add it to national income to achieve balance with the economy's expenditures.

Statistical Discrepancy As you know, it should be possible to calculate GDP either by totaling up expenditures or by summing up incomes. Either method should give the same result.

In practice, however, it is not possible for NIPA accountants to measure every input into either set of calculations with total precision. Difficulties arise due to a wide range of factors including people misreporting their incomes on tax returns and the difficulty involved with accurately estimating depreciation. As a result, the GDP number produced by the income method always differs by a small percentage from the GDP number produced by the expenditures method.

To account for this difference, NIPA accountants add a statistical discrepancy to national income. The addition of that number equalizes the GDP totals produced by the two methods. In 2012 the discrepancy value was negative \$17 billion, or less than one-half of one percent of GDP.

Table 25.3 summarizes both the expenditures approach and the income approach to GDP. The left side shows how much the U.S. economy produced in 2012 by showing how much was spent to purchase that year's output of goods and services. The right side shows how those expenditures were allocated either as income to individuals, as revenue to the government, or to other uses such as paying for the replacement of depreciated capital.

QUICK REVIEW 25.1

- Gross domestic product (GDP) is a measure of the total market value of all final goods and services produced by the economy in a specific year.
- The expenditures approach to GDP sums the total spending on final goods and services: GDP $= C + I_g + G + X_n$.
- The economy's stock of private capital expands when net investment is positive; stays constant when net investment is zero; and declines when net investment is negative.
- The income approach to GDP sums compensation to employees, rent, interest, proprietors' income, corporate profits, and taxes on production and imports to obtain national income, and then subtracts net foreign factor income and adds consumption of fixed capital and a statistical discrepancy to obtain GDP.

Other National Accounts

LO25.4 Describe the relationships among GDP, net domestic product, national income, personal income, and disposable income.

Several other national accounts provide additional useful information about the economy's performance. We can derive these accounts by making various adjustments to GDP.

Net Domestic Product

As a measure of total output, GDP does not make allowances for replacing the capital goods used up in each year's production. As a result, it does not tell us how much new output was available for consumption and for additions to the stock of capital. To determine that, we must subtract from GDP the capital that was consumed in producing the GDP and that had to be replaced. That is, we need to subtract consumption of fixed capital (depreciation) from GDP. The result is a measure of **net domestic product (NDP):**

$$\text{NDP} = \text{GDP} - \text{consumption of fixed capital (depreciation)}$$

For the United States in 2012:

	Billions
Gross domestic product	$16,245
Less: Consumption of fixed capital	2,543
Equals: Net domestic product	$13,702

NDP is simply GDP adjusted for depreciation. It measures the total annual output that the entire economy—households, businesses, government, and foreigners—can consume without impairing its capacity to produce in ensuing years.

National Income

Sometimes it is useful to know how much Americans earned for their contributions of land, labor, capital, and entrepreneurial talent. Recall that U.S. national income (NI) includes all income earned through the use of American-owned resources, whether they are located at home or abroad. It also includes taxes on production and imports. To derive NI from NDP, we must subtract the aforementioned statistical discrepancy from NDP and add net foreign factor income, since the latter is income earned by Americans overseas minus income earned by foreigners in the United States.

For the United States in 2012:

	Billions
Net domestic product	$13,702
Less: Statistical discrepancy	−17
Plus: Net foreign factor income	253
Equals: National income	$13,972

We know, too, that we can calculate national income through the income approach by simply adding up employee compensation, rent, interest, proprietors' income, corporate profit, and taxes on production and imports.

Personal Income

Personal income (PI) includes all income received, whether earned or unearned. It is likely to differ from national income (income earned) because some income earned—taxes on production and imports, Social Security taxes (payroll taxes), corporate income taxes, and undistributed corporate profits—is not received by households. Conversely, some income received—such as Social Security payments, unemployment compensation payments, welfare payments, disability and education payments to veterans, and private pension payments—is not earned. These transfer payments must be added to obtain PI.

In moving from national income to personal income, we must subtract the income that is earned but not received and add the income that is received but not earned. For the United States in 2012:

	Billions
National income	$13,972
Less: Taxes on production and imports	1,066
Less: Social Security contributions	951
Less: Corporate income taxes	435
Less: Undistributed corporate profits	542
Plus: Transfer payments	2,766*
Equals: Personal income	$13,744

*Includes statistical discrepancy and rounding error.

Disposable Income

WORKED PROBLEMS

W25.1
Measuring output and income

Disposable income (DI) is personal income less personal taxes. Personal taxes include personal income taxes, personal property taxes, and inheritance taxes. Disposable income is the amount of income that households have left over after paying their personal taxes. They are free to divide that income between consumption (C) and saving (S):

$$\text{DI} = C + S$$

For the United States in 2012:

	Billions
Personal income	$13,744
Less: Personal taxes	1,498
Equals: Disposable income	$12,246

Table 25.4 summarizes the relationships among GDP, NDP, NI, PI, and DI.

TABLE 25.4 The Relationship between GDP, NDP, NI, PI, and DI in the United States, 2012*

	Billions
Gross domestic product (GDP)	$16,245
Less: Consumption of fixed capital	2,543
Equals: Net domestic product	$13,702
Net domestic product (NDP)	$13,702
Less: Statistical discrepancy	−17
Plus: Net foreign factor income	253
Equals: National income (NI)	$13,972
National income (NI)	$13,972
Less: Taxes on production and imports	1,066
Less: Social Security contributions	951
Less: Corporate income taxes	435
Less: Undistributed corporate profits	542
Plus: Transfer payments	2,766
Equals: Personal income (PI)	$13,744
Personal income (PI)	$13,744
Less: Personal taxes	1,498
Equals: Disposable income (DI)	$12,246

*Some of the items combine categories that appear in the more detailed accounts.
Source: Bureau of Economic Analysis, **www.bea.gov**.

The Circular Flow Revisited

Figure 25.3 is an elaborate flow diagram that shows the economy's four main sectors along with the flows of expenditures and allocations that determine GDP, NDP, NI, and PI. The orange arrows represent the spending flows—$C + I_g + G + X_n$—that together measure gross domestic product. To the right of the GDP rectangle are green arrows that show first the allocations of GDP and then the adjustments needed to derive NDP, NI, PI, and DI.

The diagram illustrates the adjustments necessary to determine each of the national income accounts. For example, net domestic product is smaller than GDP because consumption of fixed capital flows away from GDP in determining NDP. Also, disposable income is smaller than personal income because personal taxes flow away from PI (to government) in deriving DI.

Note the three domestic sectors of the economy: households, government, and businesses. The household sector has an inflow of disposable income and outflows of consumption spending and savings. The government sector has an inflow of revenue in the form of types of taxes and an outflow of government disbursements in the form of purchases and transfers. The business sector has inflows from three major sources of funds for business investment and an outflow of investment expenditures.

Also, take a look at the foreign sector (all other countries) in the flow diagram. Spending by foreigners on U.S. exports adds to U.S. GDP, but some of U.S. consumption, government, and investment expenditures buy imported products. The flow from foreign markets shows that we handle this complication by calculating net exports (U.S. exports minus U.S. imports). The net export flow may be a positive or negative amount, adding to or subtracting from U.S. GDP.

Finally, you need to be aware that the flows shown in Figure 25.3 are dynamic entities and generally expand in size over time as the economy grows. But not always! Case in point: The Great Recession of 2007–2009—first discussed in the Consider This box on page 539—produced a pronounced slowing of the main spending and income flows. Specifically, U.S. businesses greatly reduced investment expenditures and households initially reduced personal consumption expenditures. Consequently, GDP, NDP, NI, and PI all significantly declined.

QUICK REVIEW 25.2

- Net domestic product (NDP) is the market value of GDP minus consumption of fixed capital (depreciation).
- National income (NI) is all income earned through the use of American-owned resources, whether located at home or abroad. NI also includes taxes on production and imports.
- Personal income (PI) is all income received by households, whether earned or not.
- Disposable income (DI) is all income received by households minus personal taxes.

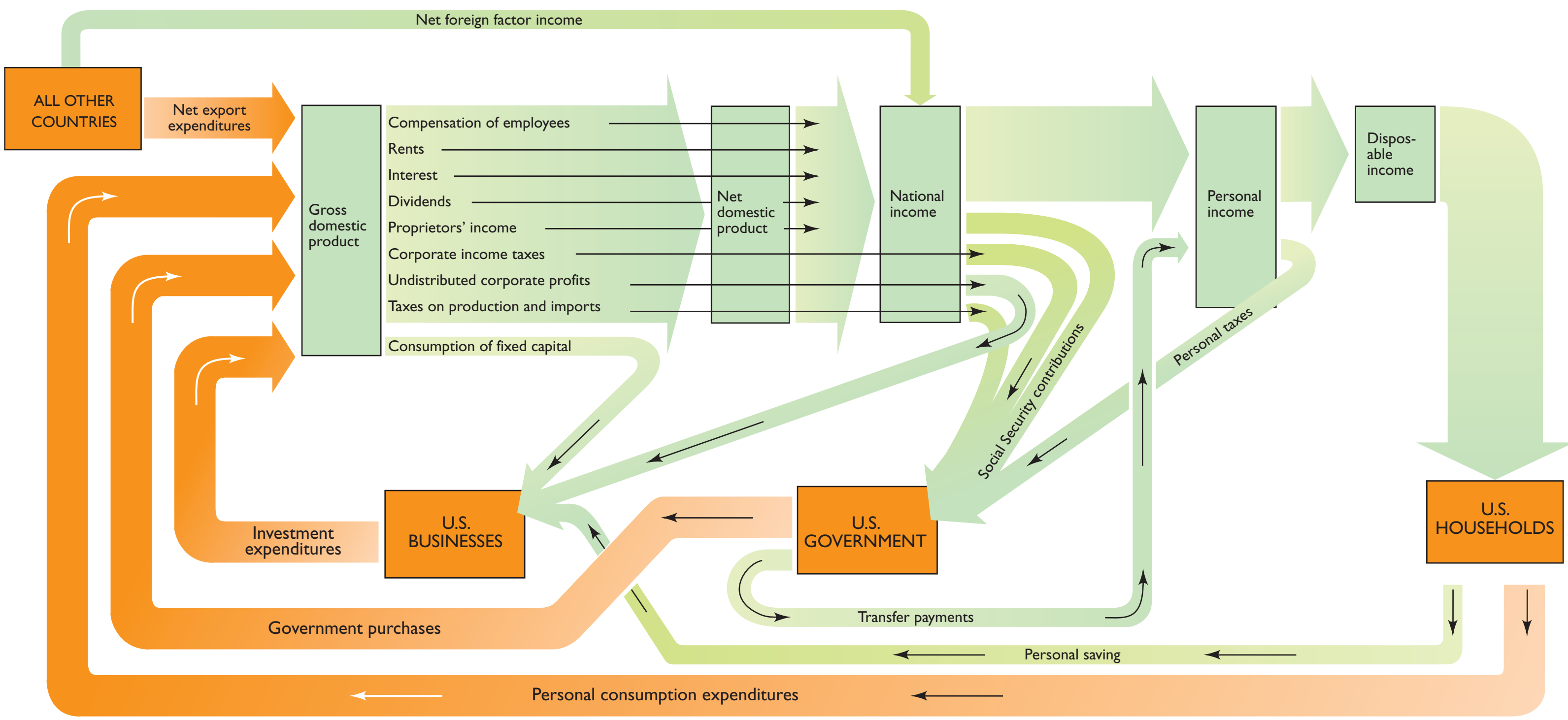
Nominal GDP versus Real GDP

LO25.5 Discuss the nature and function of a GDP price index, and describe the difference between nominal GDP and real GDP.

Recall that GDP is a measure of the market or money value of all final goods and services produced by the economy in a given year. We use money or nominal values as a common denominator to sum that heterogeneous output into a meaningful total. But, as alluded to in Chapter 24, that creates a problem: How can we compare the market values of GDP from year to year if the value of money itself changes in response to inflation (rising prices) or deflation (falling prices)? After all, we determine the value of GDP by multiplying total output by market prices.

Whether there is a 5 percent increase in output with no change in prices or a 5 percent increase in prices with no change in output, the change in the value of GDP will

FIGURE 25.3 U.S. domestic output and the flows of expenditure and income. This figure is an elaborate circular flow diagram that fits the expenditures and allocations sides of GDP to one another. The expenditures flows are shown in orange; the allocations or income flows are shown in green. You should trace through the income and expenditures flows, relating them to the five basic national income accounting measures.



be the same. And yet it is the *quantity* of goods and services that get produced and distributed to households that affects our standard of living, not the price of those goods and services. For instance, the McDonald's hamburger that sold for 95 cents in 2013 yields the same satisfaction as a nearly identical McDonald's hamburger that sold for 18 cents in 1967.

The way around this problem is to *deflate* GDP when prices rise and to *inflate* GDP when prices fall. These adjustments give us a measure of GDP for various years as if the value of the dollar had always been the same as it was in some reference year. A GDP based on the prices that prevailed when the output was produced is called unadjusted GDP, or **nominal GDP.** A GDP that has been deflated or inflated to reflect changes in the price level is called adjusted GDP, or **real GDP.**

Adjustment Process in a One-Product Economy

There are two ways we can adjust nominal GDP to reflect price changes. For simplicity, let's assume that the economy produces only one good, pizza, in the amounts indicated in Table 25.5 for years 1, 2, and 3. Suppose that we gather revenue data directly from the financial reports of the economy's pizza businesses to measure nominal GDP in various years. After completing our effort, we will have determined nominal GDP for each year, as shown in column 4 of Table 25.5. We will have no way of knowing to what extent changes in price and/or changes in quantity of output have accounted for the increases or decreases in nominal GDP that we observe.

GDP Price Index How can we determine real GDP in our pizza economy? One way is to assemble data on the price changes that occurred over various years (column 2) and use them to establish an overall price index for the entire period. Then we can use the index in each year to adjust nominal GDP to real GDP for that year.

In the actual economy, a **price index** is a measure of the price of a specified collection of goods and services, called a "market basket," in a given year as compared to the price of an identical (or highly similar) collection of goods and services in a reference year. That point of reference, or benchmark, is known as the base period, **base year,** or, simply, the reference year. More formally,

$$\begin{array}{l}\text{Price}\\\text{index}\\\text{in given}\\\text{year}\end{array} = \frac{\begin{array}{c}\text{price of market basket}\\\text{in specific year}\end{array}}{\begin{array}{c}\text{price of same market}\\\text{basket in base year}\end{array}} \times 100 \qquad (1)$$

By convention, the price ratio between a given year and the base year is multiplied by 100 to facilitate computation. For example, a price ratio of 2/1 (= 2) is expressed as a price index of 200. A price ratio of 1/3 (= 0.33) is expressed as a price index of 33.

In our pizza-only example, of course, our market basket consists of only one product. Column 2 of Table 25.5 reveals that the price of pizza was \$10 in year 1, \$20 in year 2, \$25 in year 3, and so on. Let's select year 1 as our base year. Now we can express the successive prices of the contents of our market basket in, say, years 2 and 3 as compared to the price of the market basket in year 1:

$$\text{Price index, year 2} = \frac{\$20}{\$10} \times 100 = 200$$

$$\text{Price index, year 3} = \frac{\$25}{\$10} \times 100 = 250$$

For year 1 the index has to be 100, since that year and the base year are identical.

The index numbers tell us that the price of pizza rose from year 1 to year 2 by 100 percent $\{= [(200 - 100)/100] \times 100\}$ and from year 1 to year 3 by 150 percent $\{= [(250 - 100)/100] \times 100\}$.

Dividing Nominal GDP by the Price Index We can now use the index numbers shown in column 3 to deflate

TABLE 25.5 Calculating Real GDP (Base Year = Year 1)

Year	(1) Units of Output	(2) Price of Pizza per Unit	(3) Price Index (Year 1 = 100)	(4) Unadjusted, or Nominal, GDP, (1) × (2)	(5) Adjusted, or Real, GDP
1	5	\$10	100	\$ 50	\$50
2	7	20	200	140	70
3	8	25	250	200	80
4	10	30	___	___	___
5	11	28	___	___	___

the nominal GDP figures in column 4. The simplest and most direct method of deflating is to express the index numbers as hundredths—in decimal form—and then to divide them into corresponding nominal GDP. That gives us real GDP:

$$\text{Real GDP} = \frac{\text{nominal GDP}}{\text{price index (in hundredths)}} \qquad (2)$$

WORKED PROBLEMS

W25.2

Real GDP and price indexes

Column 5 shows the results. These figures for real GDP measure the market value of the output of pizza in years 1, 2, and 3 as if the price of pizza had been a constant $10 throughout the 3-year period. In short, real GDP reveals the market value of each year's output measured in terms of dollars that have the same purchasing power as dollars had in the base year.

To test your understanding, extend Table 25.5 to years 4 and 5, using equations 1 and 2. Then run through the entire deflating procedure, using year 3 as the base period. This time you will have to inflate some of the nominal GDP data, using the same procedure as we used in the examples.

ORIGIN OF THE IDEA

O25.1

GDP price index

An Alternative Method

Another way to calculate real GDP is to gather separate data on physical outputs (as in column 1) and their prices (as in column 2) of Table 25.5. We could then determine the market value of outputs in successive years *if the base-year price ($10) had prevailed*. In year 2, the 7 units of pizza would have a value of $70 (= 7 units × $10). As column 5 confirms, that $70 worth of output is year 2's real GDP. Similarly, we could determine the real GDP for year 3 by multiplying the 8 units of output that year by the $10 price in the base year.

Once we have determined real GDP through this method, we can identify the price index for a given year simply by dividing the nominal GDP by the real GDP for that year:

$$\begin{matrix}\text{Price index} \\ \text{(in hundredths)}\end{matrix} = \frac{\text{nominal GDP}}{\text{real GDP}} \qquad (3)$$

Example: In year 2 we get a price index of 200—or, in hundredths, 2.00—which equals the nominal GDP of $140 divided by the real GDP of $70. Note that equation 3 is simply a rearrangement of equation 2. Table 25.6 summarizes the two methods of determining real GDP in our single-good economy.

TABLE 25.6 **Steps for Deriving Real GDP from Nominal GDP**

Method 1

1. Find nominal GDP for each year.
2. Compute a GDP price index.
3. Divide each year's nominal GDP by that year's price index (in hundredths) to determine real GDP.

Method 2

1. Break down nominal GDP into physical quantities of output and prices for each year.
2. Find real GDP for each year by determining the dollar amount that each year's physical output would have sold for if base-year prices had prevailed. (The GDP price index can then be found by dividing nominal GDP by real GDP.)

Real-World Considerations and Data

In the real world of many goods and services, of course, determining GDP and constructing a reliable price index are far more complex matters than in our pizza-only economy. The government accountants must assign a "weight" to each of several categories of goods and services based on the relative proportion of each category in total output. They update the weights annually as expenditure patterns change and roll the base year forward year by year using a moving average of expenditure patterns. The GDP price index used in the United States is called the *chain-type annual-weights price index*—which hints at its complexity. We spare you the details.

Table 25.7 shows some of the relationships between nominal GDP, real GDP, and the GDP price index for the

TABLE 25.7 **Nominal GDP, Real GDP, and GDP Price Index for the United States, Selected Years**

(1) Year	(2) Nominal GDP, Billions	(3) Real GDP, Billions	(4) GDP Price Index (2009 = 100)
1995	$ 7,664.0	$10,167.3	____
2000	10,289.7	____	81.9
2005	13,095.4	14,235.6	92.0
2009	14,417.9	____	100.0
2010	14,958.3	14,779.4	101.2
2012	16,244.6	15,547.0	104.5

Source: Bureau of Economic Analysis, **www.bea.gov**. All data are subject to government revision.

U.S. economy. Here the base year is 2009, where the value of the index is set at 100. Because the U.S. price level has been rising over the long run, the pre-2009 values of real GDP (column 3) are higher than the nominal values of GDP for those years (column 2). This upward adjustment acknowledges that prices were lower in the years before 2009, and thus nominal GDP understated the real output of those years in 2009 prices and must be inflated to show the correct relationship to other years.

Conversely, the rising price level of the post-2009 years caused nominal GDP figures for those years to overstate real output. So the statisticians deflate those figures to determine what real GDP would have been in other years if 2009 prices had prevailed. Doing so reveals that real GDP has been less than nominal GDP since 2009.

By inflating the nominal pre-2009 GDP data and deflating the post-2009 data, government accountants determine annual real GDP, which can then be compared with the real GDP of any other year in the series of years. So the real GDP values in column 3 are directly comparable with one another.

Once we have determined nominal GDP and real GDP, we can compute the price index. And once we have determined nominal GDP and the price index, we can calculate real GDP. Example: Nominal GDP in 2012 was \$16,244.6 billion and real GDP was \$15,547.0 billion. So the price level in 2012 was 104.5 (= \$16,244.6/\$15,547.0 × 100), or 4.5 percent higher than in 2009. If we knew the nominal GDP and the price level only, we could find the real GDP for 2012 by dividing the nominal GDP of \$15,684.8 by the 2012 price index, expressed in hundredths (1.045).

To test your understanding of the relationships between nominal GDP, real GDP, and the price level, determine the values of the price index for 1995 in Table 25.7 and determine real GDP for 2000 and 2009. We have left those figures out on purpose.

QUICK REVIEW 25.3

- Nominal GDP is output valued at current prices. Real GDP is output valued at constant base-year prices.
- The GDP price index compares the price (market value) of all the goods and services included in GDP in a given year to the price of the same market basket in a reference year.
- Nominal GDP can be transformed into real GDP by dividing the nominal GDP by the GDP price index expressed in hundredths.

Shortcomings of GDP

LO25.6 List and explain some limitations of the GDP measure.

GDP is a reasonably accurate and highly useful measure of how well or how poorly the economy is performing. But it has several shortcomings as a measure of both total output and well-being (total utility).

Nonmarket Activities

Certain productive activities do not take place in any market—the services of stay-at-home parents, for example, and the labor of carpenters who repair their own homes. Such activities never show up in GDP because the accountants who tally up GDP only get data on economic transactions involving *market activities*—that is, transactions in which output or resources are traded for money. Consequently, GDP understates a nation's total output because it does not count *unpaid work*. There is one exception: The portion of farmers' output that farmers consume themselves *is* estimated and included in GDP.

Leisure

The average workweek (excluding overtime) in the United States has declined since the beginning of the 1900s—from about 53 hours to about 35 hours. Moreover, the greater frequency of paid vacations, holidays, and leave time has shortened the work year itself. This increase in leisure time has clearly had a positive effect on overall well-being. But our system of national income accounting understates well-being by ignoring leisure's value. Nor does the system accommodate the satisfaction—the "psychic income"—that many people derive from their work.

Improved Product Quality

Because GDP is a quantitative measure rather than a qualitative measure, it fails to capture the full value of improvements in product quality. A \$200 cell phone purchased today is of very different quality than a cell phone that cost \$200 just a decade ago. Today's cell phone is digital and has greater memory capacity, a viewing screen, and quite likely a camera and a music player.

Obviously quality improvement has a great effect on economic well-being, as does the quantity of goods produced. Although the BEA adjusts GDP for quality improvement for selected items, the vast majority of such improvement for the entire range of goods and services does not get reflected in GDP.

LAST WORD

Magical Mystery Tour

The Bureau of Economic Analysis (BEA), an Agency of the Department of Commerce, Compiles the NIPA Tables. Where Does It Get the Actual Data?

Discussions of national income accounting often leave the impression that the data for the National Income and Product Accounts magically appear from some mysterious place. Let's take a tour to see where economists get their data.

Consumption The BEA derives the data for the consumption component of the GDP accounts from four main sources:

- The Census Bureau's *Retail Trade Survey*, which gains sales information from a sample of 22,000 firms.
- The Census Bureau's *Survey of Manufacturers*, which gathers information on shipments of consumer goods from 50,000 establishments.
- The Census Bureau's *Service Survey*, which collects sales data from 30,000 service businesses.
- Industry trade sources. For example, data on auto sales and aircraft are collected directly from auto and aircraft manufacturers.

Investment The sources of the data for the investment component of GDP include:

- All the sources above used to determine consumption. Purchases of capital goods are separated from purchases of consumer goods. For example, estimates of investment in equipment and software are based on manufacturers' shipments reported in the *Survey of Manufacturers*, the *Service Survey*, and industry sources.
- Census construction surveys. The Census Bureau's *Housing Starts Survey* and *Housing Sales Survey* produce the data used to measure the amount of housing construction, and the *Construction Progress Reporting Survey* is the source of data on nonresidential construction. The BEA determines changes in business inventories through the *Retail Trade Survey*, the *Wholesale Trade Survey* (of 7,100 wholesale firms), and the *Survey of Manufacturing*.

Government Purchases The data for government purchases (officially "government consumption and investment expenditures") are obtained through the following sources:

- The U.S. Office of Personnel Management, which collects data on wages and benefits, broken out by the private and public sector. Wages and benefits of government employees are the single largest "purchase" by federal, state, and local government.
- The previously mentioned Census Bureau's construction surveys, which break out private and public sector construction expenditures.
- The Census Bureau's *Survey of Government Finance*, which provides data on government consumption and investment expenditures.

Net Exports The BEA determines net exports through two main sources:

- The U.S. Customs Service, which collects data on exports and imports of goods.
- BEA surveys of potential domestic exporters and importers of services, which collect data on exports and imports of services.

So there you have it. Not so magical after all!

Source: Based on Joseph A. Ritter, "Feeding the National Accounts," Federal Reserve Bank of St. Louis *Review*, March–April 2000, pp. 11–20. For those interested, this article also provides information on the sources of data for the income side of the national accounts.

The Underground Economy

Embedded in our economy is a flourishing, productive underground sector. Some of the people who conduct business there are gamblers, smugglers, prostitutes, "fences" of stolen goods, drug growers, and drug dealers. They have good reason to conceal their incomes.

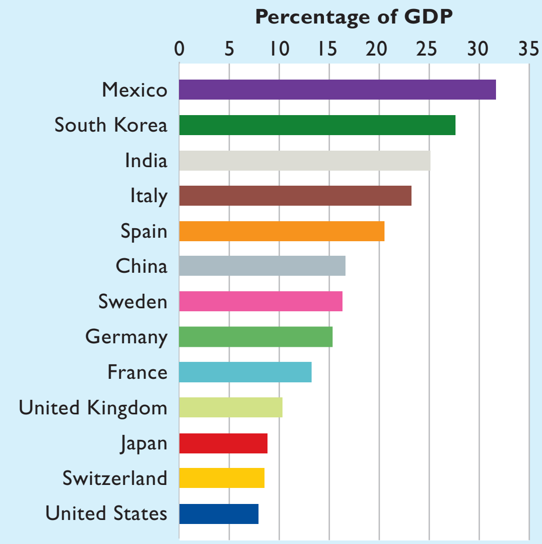
Most participants in the underground economy, however, engage in perfectly legal activities but choose illegally not to report their full incomes to the Internal Revenue Service (IRS). A barista at a coffee shop may report just a portion of the tips received from customers. Storekeepers may report only a portion of their sales receipts. Workers who want to hold on to their unemployment compensation benefits may take an "off-the-books" or "cash-only" job. A brick mason may agree to rebuild a neighbor's fireplace in exchange for the neighbor's repairing his boat engine. The value of none of these transactions shows up in GDP.

The value of underground transactions is estimated to be about 8 percent of the recorded GDP in the United States. That would mean that GDP in 2012 was understated by about $1.3 trillion. Global Perspective 25.2 shows estimates of the relative sizes of underground economies in selected nations.

GLOBAL PERSPECTIVE 25.2

The Underground Economy as a Percentage of GDP, Selected Nations

Underground economies vary in size worldwide. Three factors that help explain the variation are (1) the extent and complexity of regulation, (2) the type and degree of taxation, and (3) the effectiveness of law enforcement.



Source: Friedrich Schneider, "Shadow Economies and Corruption All Over the World: New Estimates for 145 Countries," *Economics: The Open-Access, Open-Assessment E-Journal*, vol. 1, no. 2007-9 (July 24, 2007). Used with permission of Friedrich Schneider.

GDP and the Environment

The growth of GDP is inevitably accompanied by "gross domestic by-products," including dirty air and polluted water, toxic waste, congestion, and noise. The social costs of the negative by-products reduce our economic well-being. And since those costs are not deducted from total output, GDP overstates our national well-being. Ironically, when money is spent to clean up pollution and reduce congestion, those expenses are added to GDP!

Composition and Distribution of Output

The composition of output is undoubtedly important for well-being. But GDP does not tell us whether the currently produced mix of goods and services is enriching or potentially detrimental to society. GDP assigns equal weight to an assault rifle and a set of encyclopedias, as long as both sell for the same price. Moreover, GDP reveals nothing about the way output is distributed. Does 90 percent of the output go to 10 percent of the households, for example, or is the output more evenly distributed? The distribution of output may make a big difference for society's overall well-being.

Noneconomic Sources of Well-Being

Finally, the connection between GDP and well-being is problematic for another reason. Just as a household's income does not measure its total happiness, a nation's GDP does not measure its total well-being. Many things could make a society better off without necessarily raising GDP: a reduction of crime and violence, peaceful relations with other countries, people's greater civility toward one another, better understanding between parents and children, and a reduction of drug and alcohol abuse.

QUICK REVIEW 25.4

- GDP is a reasonably accurate and very useful indicator of a nation's economic performance but should not be interpreted as a comprehensive measure of well-being.
- The major limitations of GDP as an indicator of well-being are that it fails to account for nonmarket and illegal transactions, changes in leisure and in product quality, the composition and distribution of output, and the environmental effects of production.

SUMMARY

LO25.1 Explain how gross domestic product (GDP) is defined and measured.

Gross domestic product (GDP), a basic measure of an economy's economic performance, is the market value of all final goods and services produced within the borders of a nation in a year.

Final goods are those purchased by end users, whereas intermediate goods are those purchased for resale or for further processing or manufacturing. Intermediate goods, nonproduction transactions, and secondhand sales are purposely excluded in calculating GDP.

LO25.2 Describe how expenditures on goods and services can be summed to determine GDP.

GDP may be calculated by summing total expenditures on all final output or by summing the income derived from the production of that output.

By the expenditures approach, GDP is determined by adding consumer purchases of goods and services, gross investment spending by businesses, government purchases, and net exports: $GDP = C + I_g + G + X_n$.

Personal consumption expenditures consist of expenditures on goods (durable goods and nondurable goods) and services. About 60 percent of consumer expenditures in the United States are on services, leading economists to refer to the U.S. economy as a *service economy*.

Gross investment is divided into (*a*) replacement investment (required to maintain the nation's stock of capital at its existing level) and (*b*) net investment (the net increase in the stock of capital). In most years, net investment is positive and therefore the economy's stock of capital and production capacity increase.

LO25.3 Explain how GDP can be determined by summing up all of the incomes that were derived from producing the economy's output of goods and services.

By the income or allocations approach, GDP is calculated as the sum of compensation to employees, rents, interest, proprietors' income, corporate profits, taxes on production and imports *minus* net foreign factor income, *plus* consumption of fixed capital and a statistical discrepancy.

LO25.4 Describe the relationships among GDP, net domestic product, national income, personal income, and disposable income.

Other national accounts are derived from GDP. Net domestic product (NDP) is GDP less the consumption of fixed capital. National income (NI) is total income earned by a nation's resource suppliers plus taxes on production and imports; it is found by subtracting a statistical discrepancy from NDP and adding net foreign factor income to NDP. Personal income (PI) is the total income paid to households prior to any allowance for personal taxes. Disposable income (DI) is personal income after personal taxes have been paid. DI measures the amount of income available to households to consume or save.

LO25.5 Discuss the nature and function of a GDP price index, and describe the difference between nominal GDP and real GDP.

Price indexes are computed by dividing the price of a specific collection or market basket of output in a particular period by the price of the same market basket in a base period and multiplying the result (the quotient) by 100. The GDP price index is used to adjust nominal GDP for inflation or deflation and thereby obtain real GDP.

Nominal (current-dollar) GDP measures each year's output valued in terms of the prices prevailing in that year. Real (constant-dollar) GDP measures each year's output in terms of the prices that prevailed in a selected base year. Because real GDP is adjusted for price-level changes, differences in real GDP are due only to differences in production activity.

LO25.6 List and explain some limitations of the GDP measure.

GDP is a reasonably accurate and very useful indicator of a nation's economic performance, but it has its limitations. It fails to account for nonmarket and illegal transactions, changes in leisure and in product quality, the composition and distribution of output, and the environmental effects of production. GDP should not be interpreted as a complete measure of well-being.

TERMS AND CONCEPTS

national income accounting
gross domestic product (GDP)
intermediate goods
final goods
multiple counting
value added
expenditures approach
income approach
personal consumption expenditures (C)
durable goods
nondurable goods
services
gross private domestic investment (I_g)
net private domestic investment
government purchases (G)
net exports (X_n)
taxes on production and imports
national income
consumption of fixed capital
net domestic product (NDP)
personal income (PI)
disposable income (DI)
nominal GDP
real GDP
price index
base year

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. In what ways are national income statistics useful? **LO25.1**
2. Why do national income accountants compare the market value of the total outputs in various years rather than actual physical volumes of production? What problem is posed by any comparison over time of the market values of various total outputs? How is this problem resolved? **LO25.1**
3. Which of the following goods are usually intermediate goods and which are usually final goods: running shoes, cotton fibers, watches, textbooks, coal, sunscreen lotion, lumber? **LO25.1**
4. Why do economists include only final goods and services when measuring GDP for a particular year? Why don't they include the value of the stocks and bonds bought and sold? Why don't they include the value of the used furniture bought and sold? **LO25.1**
5. Explain why an economy's output, in essence, is also its income. **LO25.1**
6. Provide three examples of each: consumer durable goods, consumer nondurable goods, and services. **LO25.2**
7. Why are changes in inventories included as part of investment spending? Suppose inventories declined by $1 billion during 2014. How would this affect the size of gross private domestic investment and gross domestic product in 2014? Explain. **LO25.2**
8. What is the difference between gross private domestic investment and net private domestic investment? If you were to determine net domestic product (NDP) through the expenditures approach, which of these two measures of investment spending would be appropriate? Explain. **LO25.2**
9. Use the concepts of gross investment and net investment to distinguish between an economy that has a rising stock of capital and one that has a falling stock of capital. Explain: "Though net investment can be positive, negative, or zero, it is impossible for gross investment to be less than zero." **LO25.2**
10. Define net exports. Explain how U.S. exports and imports each affects domestic production. How are net exports determined? Explain how net exports might be a negative amount. **LO25.2**
11. Contrast the ideas of nominal GDP and real GDP. Why is one more reliable than the other for comparing changes in the standard of living over a series of years? What is the GDP price index and what is its role in differentiating nominal GDP and real GDP? **LO25.5**
12. Which of the following are included or excluded in this year's GDP? Explain your answer in each case. **LO25.6**
    - **a.** Interest received on an AT&T corporate bond.
    - **b.** Social Security payments received by a retired factory worker.
    - **c.** Unpaid services of a family member in painting the family home.
    - **d.** Income of a dentist from the dental services provided.
    - **e.** A monthly allowance a college student receives from home.
    - **f.** Money received by Josh when he resells his nearly brand-new Honda automobile to Kim.
    - **g.** The publication and sale of a new college textbook.
    - **h.** An increase in leisure resulting from a 2-hour decrease in the length of the workweek, with no reduction in pay.
    - **i.** A $2 billion increase in business inventories.
    - **j.** The purchase of 100 shares of Google stock.
13. **LAST WORD** What government agency compiles the U.S. NIPA tables? In what U.S. department is it located? Of the several specific sources of information, name one source for each of the four components of GDP: consumption, investment, government purchases, and net exports.

REVIEW QUESTIONS

1. Tina walks into Ted's sporting goods store and buys a punching bag for $100. That $100 payment counts as ____________ for Tina and ___________ for Ted. **LO25.1**
    - **a.** Income; expenditure.
    - **b.** Value added; multiple counting.
    - **c.** Expenditure; income.
    - **d.** Rents; profits.
2. Which of the following transactions would count in GDP? **LO25.1**

    *Select **one or more** answers from the choices shown.*
    - **a.** Kerry buys a new sweater to wear this winter.
    - **b.** Patricia receives a Social Security check.
    - **c.** Roberto gives his daughter $50 for her birthday.
    - **d.** Latika sells $1,000 of General Electric stock.
    - **e.** Karen buys a new car.
    - **f.** Amy buys a used car.
3. A small economy starts the year with $1 million in capital. During the course of the year, gross investment is $150,000 and depreciation is $50,000. How big is the economy's stock of capital at the end of the year? **LO25.2**
    - **a.** $1,150,000.
    - **b.** $1,100,000.
    - **c.** $1,000,000.
    - **d.** $850,000.
    - **e.** $800,000.
4. Suppose that this year a small country has a GDP of $100 billion. Also assume that I_g = $30 billion, C = $60 billion, and X_n = − $10 billion. How big is G? **LO25.3**
    - **a.** $0.
    - **b.** $10 billion.
    - **c.** $20 billion.
    - **d.** $30 billion.

5. Suppose that California imposes a sales tax of 10 percent on all goods and services. A Californian named Ralph then goes into a home improvement store in the state capital of Sacramento and buys a leaf blower that is priced at $200. With the 10 percent sales tax, his total comes to $220. How much of the $220 paid by Ralph will be counted in the national income and product accounts as private income (employee compensation, rents, interest, proprietor's income, and corporate profits)? **LO25.3**
   a. $220.
   b. $200.
   c. $180.
   d. None of the above.
6. Suppose GDP is $16 trillion, with $10 trillion coming from consumption, $2 trillion coming from gross investment, $3.5 trillion coming from government expenditures, and $500 billion coming from net exports. Also suppose that across the whole economy, depreciation (consumption of fixed capital) totals $1 trillion. From these figures, we see that net domestic product equals: **LO25.4**
   a. $17.0 trillion.
   b. $16.0 trillion.
   c. $15.5 trillion.
   d. None of the above.
7. Suppose GDP is $15 trillion, with $8 trillion coming from consumption, $2.5 trillion coming from gross investment, $3.5 trillion coming from government expenditures, and $1 trillion coming from net exports. Also suppose that across the whole economy, personal income is $12 trillion. If the government collects $1.5 trillion in personal taxes, then disposable income will be: **LO25.4**
   a. $13.5 trillion.
   b. $12.0 trillion.
   c. $10.5 trillion.
   d. None of the above.
8. Suppose that this year's nominal GDP is $16 trillion. To account for the effects of inflation, we construct a price-level index in which an index value of 100 represents the price level five years ago. Using that index, we find that this year's real GDP is $15 trillion. Given those numbers, we can conclude that the current value of the index is: **LO25.5**
   a. Higher than 100.
   b. Lower than 100.
   c. Still 100.
9. Which of the following items will be included in official U.S. GDP statistics? **LO25.6**
   *Select **one or more** answers from the choices shown.*
   a. Revenue generated by illegal marijuana growers in Oregon.
   b. Money spent to clean up a local toxic waste site in Ohio.
   c. Revenue generated by legal medical marijuana sales in California.
   d. The dollar value of the annoyance felt by local citizens living near a noisy airport in Georgia.
   e. Robert paying Ted for a haircut in Chicago.
   f. Emily and Rhonda trading an hour of dance lessons for a haircut in Dallas.

PROBLEMS

1. Suppose that annual output in year 1 in a 3-good economy is 3 quarts of ice cream, 1 bottle of shampoo, and 3 jars of peanut butter. In year 2, the output mix changes to 5 quarts of ice cream, 2 bottles of shampoo, and 2 jars of peanut butter. If the prices in both years are $4 per quart for ice cream, $3 per bottle of shampoo, and $2 per jar of peanut butter, what was the economy's GDP in year 1? What was its GDP in year 2? **LO25.1**
2. Assume that a grower of flower bulbs sells its annual output of bulbs to an Internet retailer for $70,000. The retailer, in turn, brings in $160,000 from selling the bulbs directly to final customers. What amount would these two transactions add to personal consumption expenditures and thus to GDP during the year? **LO25.1**
3. If in some country personal consumption expenditures in a specific year are $50 billion, purchases of stocks and bonds are $30 billion, net exports are −$10 billion, government purchases are $20 billion, sales of secondhand items are $8 billion, and gross investment is $25 billion, what is the country's GDP for the year? **LO25.2**
4. To the right is a list of domestic output and national income figures for a certain year. All figures are in billions. The questions that follow ask you to determine the major national

Personal consumption expenditures	$245
Net foreign factor income	4
Transfer payments	12
Rents	14
Consumption of fixed capital (depreciation)	27
Statistical discrepancy	8
Social Security contributions	20
Interest	13
Proprietors' income	33
Net exports	11
Dividends	16
Compensation of employees	223
Taxes on production and imports	18
Undistributed corporate profits	21
Personal taxes	26
Corporate income taxes	19
Corporate profits	56
Government purchases	72
Net private domestic investment	33
Personal saving	20

income measures by both the expenditures and the income approaches. The results you obtain with the different methods should be the same. **LO25.4**

a. Using the above data, determine GDP by both the expenditures and the income approaches. Then determine NDP.

b. Now determine NI in two ways: first, by making the required additions or subtractions from NDP; and second, by adding up the types of income and taxes that make up NI.

c. Adjust NI (from part *b*) as required to obtain PI.

d. Adjust PI (from part *c*) as required to obtain DI.

5. Using the following national income accounting data, compute (*a*) GDP, (*b*) NDP, and (*c*) NI. All figures are in billions. **LO25.4**

Compensation of employees	$194.2
U.S. exports of goods and services	17.8
Consumption of fixed capital	11.8
Government purchases	59.4
Taxes on production and imports	14.4
Net private domestic investment	52.1
Transfer payments	13.9
U.S. imports of goods and services	16.5
Personal taxes	40.5
Net foreign factor income	2.2
Personal consumption expenditures	219.1
Statistical discrepancy	0

6. Suppose that in 1984 the total output in a single-good economy was 7,000 buckets of chicken. Also suppose that in 1984 each bucket of chicken was priced at $10. Finally, assume that in 2005 the price per bucket of chicken was $16 and that 22,000 buckets were produced. Determine the GDP price index for 1984, using 2005 as the base year. By what percentage did the price level, as measured by this index, rise between 1984 and 2005? What were the amounts of real GDP in 1984 and 2005? **LO25.5**

7. The following table shows nominal GDP and an appropriate price index for a group of selected years. Compute real GDP. Indicate in each calculation whether you are inflating or deflating the nominal GDP data. **LO25.5**

Year	Nominal GDP, Billions	Price Index (2005 = 100)	Real GDP, Billions
1968	$ 909.8	22.01	$_____
1978	2,293.8	40.40	$_____
1988	5,100.4	66.98	$_____
1998	8,793.5	85.51	$_____
2008	14,441.4	108.48	$_____

8. Assume that the total value of the following items is $600 billion in a specific year for Upper Mongoose: net exports = $50 billion; value of new goods and services produced in the underground economy = $75 billion; personal consumption expenditures = $300 billion; value of the services of stay-at-home parents = $25 billion; gross domestic investment = $100 billion; government purchases = $50 billion. What is Upper Mongoose's GDP for the year? What is the size of the underground economy as a percentage of GDP? By what percentage would GDP be boosted if the value of the services of stay-at-home spouses were included in GDP? **LO25.6**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **26**

Economic Growth

Learning Objectives

LO26.1 List two ways that economic growth is measured.

LO26.2 Define "modern economic growth" and explain the institutional structures needed for an economy to experience it.

LO26.3 Identify the general supply, demand, and efficiency forces that give rise to economic growth.

LO26.4 Describe "growth accounting" and the specific factors accounting for economic growth in the United States.

LO26.5 Explain why the trend rate of U.S. productivity growth has increased since the earlier 1973–1995 period.

LO26.6 Discuss differing perspectives as to whether growth is desirable and sustainable.

People living in rich countries tend to take economic growth and rising standards of living for granted. Recessions—periods during which output declines—are normally infrequent and temporary, usually lasting less than a year. Once they pass, modern capitalistic economies return to growing, and living standards continue their seemingly inexorable rise.

But a look back at history or a look around the world today quickly dispels any confidence that economic growth and rising standards of living are automatic or routine. Historically, continually rising living standards are a recent phenomenon, seen only during the last century or two. Before that time, living standards barely rose—if at all—from one generation to the next. And a look around the world today reveals huge differences in standards of living resulting from the disturbing fact that, although some countries have enjoyed decades or even centuries of steadily rising per capita income levels, other countries have experienced hardly any economic growth at all.

This chapter investigates the causes of economic growth, what institutional structures appear to promote economic growth, and the controversies surrounding the benefits and costs of economic growth. As you will see, economic growth has been perhaps the most revolutionary and powerful force in history. Consequently, no study of economics is complete without a thorough understanding of the causes and consequences of economic growth.

Economic Growth

LO26.1 List two ways that economic growth is measured.

Economists define and measure **economic growth** as either:

- An increase in real GDP occurring over some time period.
- An increase in real GDP per capita occurring over some time period.

With either definition, economic growth is calculated as a percentage rate of growth per quarter (3-month period) or per year. For the first definition, for example, real GDP in the United States was \$15,052.4 billion in 2011 and \$15,470.0 in 2012. So the U.S. economic growth rate for 2007 was 2.8 percent {= [(15,470.0 billion − \$15,052.4 billion)/\$15,052.4 billion] × 100}. Growth rates normally are positive, but not always. In recession year 2009, for instance, the U.S. rate of economic growth was a *minus* 2.4 percent.

The second definition of economic growth in the bulleted list takes into consideration the size of the population. **Real GDP per capita** (or per capita output) is the amount of real output per person in a country. It is calculated, as follows.

$$\text{Real GDP per capita} = \frac{\text{Real GDP}}{\text{Population}}$$

For example, in 2011 the real GDP in the United States was \$15,052.4 billion and population was 311.6 million. Therefore, real GDP per capita in that year was \$48,307. In 2012 real GDP per capita increased to \$49,283. So the growth rate of real GDP per capita in 2012 was 2.0 percent {= [(\$49,283 − \$48,307)/\$48,307] × 100}. In contrast, real GDP per capita fell by 3.3 percent in recession year 2009.

For measuring expansion of military potential or political preeminence, the growth of real GDP is more useful. Unless specified otherwise, growth rates reported in the news and by international agencies use this definition of economic growth. For comparing living standards, however, the second definition is superior. While China's GDP in 2012 was \$12,380 billion compared with Denmark's \$332 billion, Denmark's real GDP per capita was \$37,700 compared with China's hugely lower \$9,100. And in some cases growth of real GDP can be misleading. The African nation of Eritrea had real GDP growth of 1.3 percent per year from 2000–2008. But over the same period its annual growth of population was 3.8 percent, resulting in a decline in real GDP per capita of roughly 2.5 percent per year.

Growth as a Goal

Growth is a widely held economic goal. The expansion of total output relative to population results in rising real wages and incomes and thus higher standards of living. An economy that is experiencing economic growth is better able to meet people's wants and resolve socioeconomic problems. Rising real wages and income provide richer opportunities to individuals and families—a vacation trip, a personal computer, a higher education—without sacrificing other opportunities and pleasures. A growing economy can undertake new programs to alleviate poverty, embrace diversity, cultivate the arts, and protect the environment without impairing existing levels of consumption, investment, and public goods production.

In short, *growth lessens the burden of scarcity*. A growing economy, unlike a static economy, can consume more today while increasing its capacity to produce more in the future. By easing the burden of scarcity—by relaxing society's constraints on production—economic growth enables a nation to attain its economic goals more readily and to undertake new endeavors that require the use of goods and services to be accomplished.

Arithmetic of Growth

Why do economists pay so much attention to small changes in the rate of economic growth? Because those changes really matter! For the United States, with a current nominal GDP of about \$16.2 trillion, the difference between a 3 percent and a 4 percent rate of growth is about $162 billion of output each year. For a poor country, a difference of one-half of a percentage point in the rate of growth may mean the difference between starvation and mere hunger.

The mathematical approximation called the **rule of 70** provides a quantitative grasp of the effect of economic growth. The rule of 70 tells us that we can find the number of years it will take for some measure to double, given its annual percentage increase, by dividing that percentage increase into the number 70. So

$$\begin{array}{c}\text{Approximate}\\ \text{number of years}\\ \text{required to double}\\ \text{real GDP}\end{array} = \frac{70}{\begin{array}{c}\text{annual percentage rate}\\ \text{of growth}\end{array}}$$

Examples: A 3 percent annual rate of growth will double real GDP in about 23 (= 70 ÷ 3) years. Growth of 8 percent per year will double real GDP in about 9 (= 70 ÷ 8) years. The rule of 70 is applicable generally. For example, it works for estimating how long it will take the price level or a savings account to double at various percentage rates of inflation or interest. When compounded over many years, an apparently small difference in the rate of growth thus becomes highly significant. Suppose China and Italy start with identical GDPs, but then China grows at an 8 percent yearly rate, while Italy grows at 2 percent. China's GDP would double in about 9 years, while Italy's GDP would double in 35 years.

WORKED PROBLEMS

W26.1
GDP growth

Growth in the United States

Table 26.1 gives an overview of economic growth in the United States since 1950. Column 2 reveals strong growth as measured by increases in real GDP. Note that between 1950 and 2012 real GDP increased more than sevenfold. But the U.S. population also increased. Nevertheless, in column 4 we find that real GDP per capita rose more than threefold over these years.

TABLE 26.1 Real GDP and Real GDP per Capita, Selected Years, 1950–2012

(1) Year	(2) Real GDP, Billions of 2009 $	(3) Population, Millions	(4) Real GDP Per Capita, 2009 $ (2) ÷ (3)
1950	$ 2,182	152	$14,355
1960	3,106	181	17,160
1970	4,718	205	23,015
1980	6,443	228	28,259
1990	8,945	250	35,780
2000	12,565	282	44,557
2012	15,471	313	49,428

Source: Data are from the Bureau of Economic Analysis, **www.bea.gov**, and the U.S. Census Bureau, **www.census.gov**. All data are subject to government revision.

What has been the *rate* of U.S. growth? Real GDP grew at an annual rate of about 3.2 percent between 1950 and 2012. Real GDP per capita increased at roughly 2 percent per year over that time. But we must qualify these raw numbers in several ways:

- ***Improved products and services*** Since the numbers in Table 26.1 do not fully account for improvements in products and services, they understate the growth of economic well-being. Such purely quantitative data do not fully compare an era of vacuum tube computers and low-efficiency V8 hot rods with an era of digital cell phone networks and fuel-sipping, hybrid-drive vehicles.
- ***Added leisure*** The increases in real GDP and per capita GDP identified in Table 26.1 were accomplished despite increases in leisure. The average workweek, once 50 hours, is now about 35 hours (excluding overtime hours). Again the raw growth numbers understate the gain in economic well-being.
- ***Other impacts*** These measures of growth do not account for any effects growth may have had on the environment and the quality of life. If growth debases the physical environment, excessively warms the planet, and creates a stressful work environment, the bare growth numbers will overstate the gains in well-being that result from growth. On the other hand, if growth leads to stronger environmental protections or a more secure and stress-free lifestyle, these numbers will understate the gains in well-being.

In Chapter 24, we made two other key points about U.S. growth rates. First, they are not constant or smooth over time. Like those of other countries, U.S. growth rates vary quarterly and annually depending on a variety of factors such as the introduction of major new inventions and the economy's current position in the business cycle. Second, many countries share the U.S. experience of positive and ongoing economic growth. But sustained growth is both a historically new occurrence and also one that is not shared equally by all countries.

QUICK REVIEW 26.1

- Economists measure economic growth as either (a) an increase in real GDP over time or (b) an increase in real GDP per capita over time.
- Real GDP in the United States has grown at an average annual rate of about 3.2 percent since 1950; real GDP per capita has grown at roughly a 2 percent annual rate over that same period.

Modern Economic Growth

LO26.2 Define "modern economic growth" and explain the institutional structures needed for an economy to experience it.

We now live in an era of wireless high-speed Internet connections, genetic engineering, and space exploration. New inventions and new technologies drive continual economic growth and ongoing increases in living standards. But it wasn't always like this. Economic growth and sustained increases in living standards are a historically recent phenomenon that started with the Industrial Revolution of the late 1700s. Before the Industrial Revolution, living standards were basically flat over long periods of time so that, for instance, Greek peasants living in the year 300 B.C. had about the same material standard of living as Greek peasants living in the year A.D. 1500. By contrast, our current era of **modern economic growth** is characterized by sustained and ongoing increases in living standards that can cause dramatic increases in the standard of living within less than a single human lifetime.

Economic historians informally date the start of the Industrial Revolution to the year 1776, when the Scottish inventor James Watt perfected a powerful and efficient steam engine. This steam engine inaugurated the modern era since the device could be used to drive industrial factory equipment, steamships, and steam locomotives.

The new industrial factories mass-produced goods for the first time. This meant that nearly all manufacturing shifted from items produced by hand by local craftsmen to items mass-produced in distant factories. The new steamships and steam locomotives meant that resources could easily flow to factories and that the products of factories could be shipped to distant consumers at low cost. The result was a huge increase in long-distance trade and a major population shift as people left farms to go work in the towns and cities where the new industrial factories were concentrated.

Steam power would later be largely replaced by electric power, and many more inventions would follow the steam engine that started the Industrial Revolution. These included railroads, motorized vehicles, telephones, airplanes, container ships, computers, the Internet, and many more. But the key point is that the last 200 or so years of history have been fundamentally different from anything that went before.

The biggest change has been change itself. Whereas in earlier times material standards of living and the goods and services that people produced and consumed changed very little even over the course of an entire human life span, today people living in countries experiencing modern economic growth are constantly exposed to new technologies, new products, and new services.

What is more, modern economic growth has vastly affected cultural, social, and political arrangements.

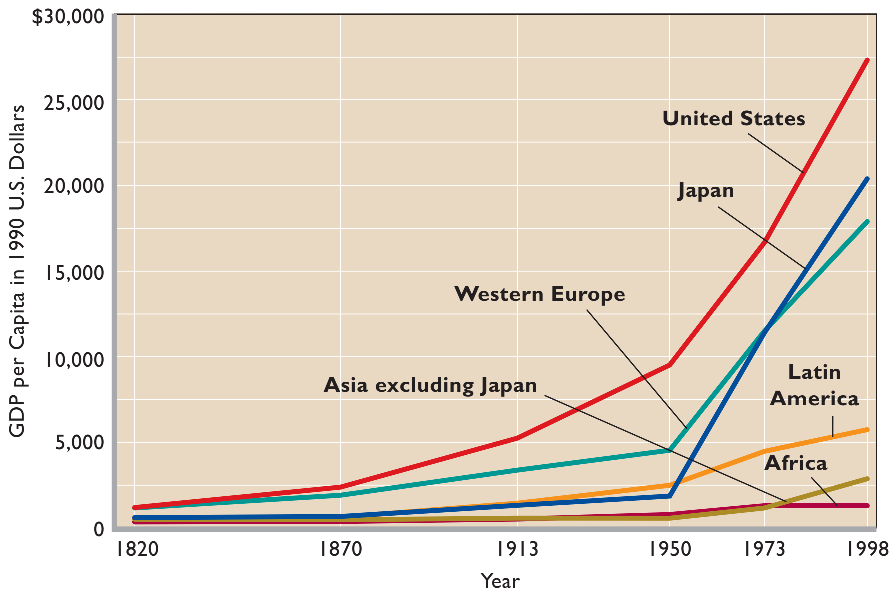
- Culturally, the vast increases in wealth and living standards have allowed ordinary people for the first time in history to have significant time for leisure activities and the arts.
- Socially, countries experiencing modern economic growth have abolished feudalism, instituted universal public education, and largely eliminated ancient social norms and legal restrictions against women and minorities doing certain jobs or holding certain positions.
- Politically, countries experiencing modern economic growth have tended to move toward democracy, a form of government that was extremely rare before the start of the Industrial Revolution.

In addition, the average human lifespan has more than doubled, from an average of less than 30 years before modern economic growth began in the late 1700s to a worldwide average of over 67 years today. Thus, for the first time in world history, the average person can expect to live into old age. These and other changes speak to the truly revolutionary power of economic growth and naturally lead economists to consider the causes of economic growth and what policies could be pursued to sustain and promote it. Their desire is intensified by the reality that economic growth is distributed so unevenly around the world.

The Uneven Distribution of Growth

Modern economic growth has spread only slowly from its British birthplace. It first advanced to France, Germany, and other parts of western Europe in the early 1800s before spreading to the United States, Canada, and Australia by the mid 1800s. Japan began to industrialize in the 1870s, but the rest of Asia did not follow until the early to mid 1900s, at which time large parts of Central and South America as well

FIGURE 26.1 The great divergence in standards of living. Income levels around the world were very similar in 1820. But they are now very different because certain areas, including the United States and western Europe, began experiencing modern economic growth much earlier than other areas.



Source: Angus Maddison, *The World Economy: A Millennial Perspective* (Paris: OECD, 2001), p. 264.

as the Middle East also began to experience modern economic growth. Most recent has been Africa, which for the most part did not experience modern economic growth until the last few decades. Notably, some parts of the world have yet to experience modern economic growth at all.

The different starting dates for modern economic growth in various parts of the world are the main cause of the vast differences in per capita GDP levels seen today. The current huge gaps between rich countries like the United States and Japan and poor countries like North Korea and Burundi were shown previously in Global Perspective 24.1. But the huge divergence in living standards caused by the fact that different countries started modern economic growth at different times is best seen in Figure 26.1, which shows how GDP per capita has evolved since 1820 in the United States, western Europe, Latin America, Asia, and Africa.

To make the comparison of living standards easier, income levels in all places and at all times have been converted into 1990 U.S. dollars. Using this convention, it is clear that in 1820 per capita incomes in all areas were quite similar, with the richest area in the world in 1820, western Europe, having an average per capita income of $1,232, while the poorest area of the world at that time, Africa, had an average per capita income of $418. Thus, in 1820, average incomes in the richest area were only about three times larger than those in the poorest area.

But because western Europe and the United States started experiencing modern economic growth earlier than other areas, they have now ended up vastly richer than other areas, despite the fact that per capita incomes in nearly all places have increased at least a bit. For instance, per capita GDP in the United States in 1998 was $27,331 while it was only $1,368 in Africa. Thus, because modern economic growth has occurred for nearly two centuries in the United States compared to a few decades in Africa, average living standards in the United States in 1998 were nearly 20 times higher than those in Africa.

Catching Up Is Possible

Do not get the wrong impression looking at Figure 26.1. Countries that began modern economic growth more recently are *not* doomed to be permanently poorer than the countries that began modern economic growth at an earlier date. This is true because people can adopt technology more quickly than they can invent it. Broadly speaking, the richest countries today have achieved that status because they have the most advanced technology. But because they already have the most advanced technology, they must invent new technology to get even richer. Because inventing and implementing new technology is slow and costly, real GDP per capita in the richest **leader countries** typically grows by an average annual rate of just 2 or 3 percent per year.

By contrast, poorer **follower countries** can grow much faster because they can simply adopt existing technologies from rich leader countries. For instance, in many places in Africa today, the first telephones most people have ever been able to use are cell phones. That is, these countries have not even bothered to install the copper wires necessary for land-line telephones, which are basically a nineteenth-century technology. Instead, they have gone directly for Internet-capable mobile phone networks, a twenty-first-century technology. By doing so, they skip past many stages of technology and development that the United States and other currently rich countries had to pass through. In effect, they jump directly to the most modern, most highly productive technology. The result is that, under the right circumstances, it is possible for poorer countries to experience extremely rapid increases in living standards. This can continue until they have caught up with the leader countries and become leader countries themselves. Once that happens, their growth rates fall down to the 2 or 3 percent rate typical of leader countries. This happens because once they are also rich and using the latest technology, their growth rates are limited by the rate at which new technology can be invented and applied.

Table 26.2 shows both how the growth rates of leader countries are constrained by the rate of technological progress as well as how certain follower countries have been able to catch up by adopting more advanced technologies and growing rapidly. Table 26.2 shows real GDP per capita in 1960 and 2010 as well as the average annual growth rate of real GDP per capita between 1960 and 2010 for three countries—the United States, the United Kingdom, and France—that were already rich leader countries in 1960 as well as for five other nations that were relatively poor follower countries at that time. To make comparisons easy, the GDPs and GDPs per capita for all countries are expressed in terms of 2005 U.S. dollars. The countries are ordered by their respective GDPs per capita in 1960, so that the richest country in the world at the time, the United States, is listed first while the poorest of the eight selected countries at the time, South Korea, is listed last.

First, notice that the average annual growth rates of the three leader countries—the United States, the United Kingdom, and France—have all been between 2.1 and 2.5 percent per year because their growth rates are limited by the rate at which new technologies can be invented and applied. By contrast, the five countries that were follower countries in 1960 have been able to grow much faster, between 3.3 percent per year and 5.4 percent per year. This has had remarkable effects on their standards of living relative to the leader countries. For instance, Ireland's

TABLE 26.2 Real GDP per Capita in 1960 and 2010 Plus Average Annual Growth Rates of Real GDP per Capita from 1960–2010 for Selected Countries. (Figures are in 2005 dollars.)

Country	Real GDP per Capita, 1960	Real GDP per Capita, 2010	Average Annual Growth Rate, 1960–2010
United States	$14,766	$41,365	2.1
United Kingdom	11,257	34,268	2.2
France	9,347	31,299	2.4
Ireland	6,666	34,877	3.3
Japan	5,472	31,477	3.5
Singapore	4,149	55,862	5.2
Hong Kong	3,849	38,865	4.6
South Korea	1,765	26,609	5.4

Note: GDP figures for all countries are measured in "international dollars" of equal value to U.S. dollars in 2005.

Source: Penn World Table version 6.3, **pwt.econ.upenn.edu**. Used by permission of the Center for International Comparisons at the University of Pennsylvania.

CONSIDER THIS . . .

Economic Growth Rates Matter!

When compounded over many decades, small absolute differences in rates of economic growth add up to substantial differences in real GDP and standards of living. Consider three hypothetical countries—Slogo, Sumgo, and Speedo. Suppose that in 2014 these countries have identical levels of real GDP ($6 trillion), population (200 million), and real GDP per capita ($30,000). Also, assume that annual real GDP growth is 2 percent in Slogo, 3 percent in Sumgo, and 4 percent in Speedo.

How will these alternative growth rates affect real GDP and real GDP per capita over a long period, say, a 70-year lifespan? By 2084 the 2, 3, and 4 percent growth rates would boost real GDP from $6 trillion to:

- $24 trillion in Slogo.
- $47 trillion in Sumgo.
- $93 trillion in Speedo.

For illustration, let's assume that each country experienced an average annual population growth of 1 percent over the 70 years. Then, in 2084 real GDP per capita would be about:

- $60,000 in Slogo.
- $118,000 in Sumgo.
- $233,000 in Speedo.

Even small differences in growth rates matter!

GDP per capita was only 60 percent that of its neighbor, the United Kingdom, in 1960. But because Ireland grew at a 3.3 percent rate for the next 50 years while the United Kingdom grew at only a 2.2 percent rate over that time period, by 2010 Ireland's GDP per capita was actually higher than the United Kingdom's GDP per capita. Ireland had become a leader country, too.

The growth experiences of the other four nations that were poor in 1960 have been even more dramatic. Hong Kong, for instance, moved from a GDP per capita that was less than one-third of that enjoyed by the United Kingdom in 1960 to a GDP per capita 13 percent higher than that of the United Kingdom in 2010. The Consider This box on the previous page emphasizes both how quickly small differences in growth rates can change the level of real GDP per capita and how countries stand in relation to each other in terms of real GDP per capita.

Finally, you may be puzzled as to why the GDP per capita of the United States in 2011 in Table 26.2 is so much higher than that of other rich leader countries. Why, for instance, is U.S. GDP per capita 32 percent higher than French GDP per capita? One important reason is that U.S. citizens put in substantially more labor time than do the citizens of most other leader countries. First, a much larger fraction of the U.S. population is employed than in other rich leader countries. Second, U.S. employees work many more hours per year than do employees in other rich leader countries. For example, 58 percent of the working-age population of the United States was employed in 2010 compared to 51 percent in France. That's a difference of about 14 percent. And American employees worked an average of 1,778 total hours during 2010, compared to an average of 1,478 total hours for French workers. That's a difference of about 20 percent. Added together, these two differences between U.S. and French labor supply imply about a 34 percent difference in the total number of hours worked in the French and American economies. Thus, differences in labor supply help explain differences between rich leader countries in terms of their differing levels of GDP per person.

Buy why do Americans supply so much more labor than workers in France and some of the other rich leader countries? Explanations put forth by economists include cultural differences regarding the proper balance between work and leisure, stronger unions in France and other rich leader countries, and more generous unemployment and welfare programs in France and other rich leader countries. France and other rich leader countries also tend to have higher tax rates than the United States—something that may significantly discourage employment. And, finally, the legal workweek is shorter in some countries than it is in the United States.

QUICK REVIEW 26.2

- Before the advent of modern economic growth starting in England in the late 1700s, living standards showed no sustained increases over time.
- Large differences in standards of living exist today because certain areas like the United States have experienced nearly 200 years of modern economic growth while other areas have had only a few decades of economic growth.
- Poor follower countries can catch up with and even surpass the living standards of rich leader countries by adopting the cutting-edge technologies and institutions already developed by rich leader countries.
- Substantial differences in GDP per capita among technologically advanced leader countries are often caused by differences in the amount of labor supplied.

Institutional Structures That Promote Modern Economic Growth

Table 26.2 demonstrates that poor follower countries can catch up and become rich leader countries by growing rapidly. But how does a country start that process and enter into modern economic growth? And once it has started modern economic growth, how does it keep the process going?

Economic historians have identified several institutional structures that promote and sustain modern economic growth. Some structures increase the savings and investment that are needed to fund the construction and maintenance of the huge amounts of infrastructure required to run modern economies. Other institutional structures promote the development of new technologies. And still others act to ensure that resources flow efficiently to their most productive uses. These growth-promoting institutional structures include:

- ***Strong property rights*** These appear to be absolutely necessary for rapid and sustained economic growth. People will not invest if they believe that thieves, bandits, or a rapacious and tyrannical government will steal their investments or their expected returns.
- ***Patents and copyrights*** Before patents and copyrights were first issued and enforced, inventors and authors usually saw their ideas stolen before they could profit from them. By giving inventors and authors the exclusive right to market and sell their creations, patents and copyrights give a strong financial incentive to invent and create.
- ***Efficient financial institutions*** These are needed to channel the savings generated by households toward the businesses, entrepreneurs, and inventors that do

most of society's investing and inventing. Banks as well as stock and bond markets appear to be institutions crucial to modern economic growth.

- ***Literacy and widespread education*** Without highly educated inventors, new technologies do not get developed. And without a highly educated workforce, it is impossible to implement those technologies and put them to productive use.
- ***Free trade*** Free trade promotes economic growth by allowing countries to specialize so that different types of output can be produced in the countries where they can be made at the lowest opportunity cost. In addition, free trade promotes the rapid spread of new ideas so that innovations made in one country quickly spread to other countries.
- ***A competitive market system*** Under a market system, prices and profits serve as the signals that tell firms what to make and how much of it to make. Rich leader countries vary substantially in terms of how much government regulation they impose on markets, but in all cases, firms have substantial autonomy to follow market signals in deciding on current production and in making investments to produce what they believe consumers will demand in the future.

Several other difficult-to-measure factors also influence a nation's capacity for economic growth. The overall social-cultural-political environment of the United States, for example, has encouraged economic growth. Beyond the market system that has prevailed in the United States, the United States also has had a stable political system characterized by democratic principles, internal order, the right of property ownership, the legal status of enterprise, and the enforcement of contracts. Economic freedom and political freedom have been "growth-friendly."

In addition, and unlike some nations, there are virtually no social or moral taboos on production and material progress in the United States. The nation's social philosophy has embraced wealth creation as an attainable and desirable goal and the inventor, the innovator, and the businessperson are accorded high degrees of prestige and respect in American society. Finally, Americans have a positive attitude toward work and risk taking, resulting in an ample supply of willing workers and innovative entrepreneurs. A flow of energetic immigrants has greatly augmented that supply.

The nearby Consider This box deals with how fast-growing follower countries such as India sometimes alter their growth-related institutional structures as they grow richer. Web Chapter 39 looks at the special problems of economic growth in developing nations.

CONSIDER THIS . . .

Patents and Innovation

It costs U.S. and European drug companies about $1 billion to research, patent, and safety-test a new drug because literally thousands of candidate drugs fail for each drug that succeeds. The only way to cover these costs is by relying on patent protections that give a drug's developer the exclusive monopoly right to market and sell the new drug for 20 years following the patent application. The revenues over that time period will hopefully be enough to cover the drug's development costs and—if the drug is popular—generate a profit for the drug company.

Leader and follower countries have gotten into heated disputes over patented drugs, however, because the follower countries have often refused to recognize the patents granted to pharmaceutical companies in rich countries. India, for instance, has allowed local drug companies to copy and sell drugs that were developed by U.S. companies and are still under patent protection in the United States.

That policy benefits Indian consumers because competition among the local drug companies drives down the price to below the monopoly price that would be charged by the patent owner. But the weak patent protections in India have a side effect. They make it completely unprofitable for local drug producers to try to develop innovative new drugs. Local rivals would simply copy the new drugs and sell them at very low prices. So India has recently moved to strengthen its patent protections to try to provide financial incentives to transform its local drug companies from copycats into innovators. But note that the innovative new drugs that may result from the increased patent protections are not without a cost. As patent protections in India are improved, inexpensive local drugs copied from the leader countries will no longer be available to Indian consumers.

Determinants of Growth

LO26.3 Identify the general supply, demand, and efficiency forces that give rise to economic growth.

Our discussion of modern economic growth and the institutional structures that promote it has purposely been general. We now want to focus our discussion on six factors that directly affect the *rate* and quality of economic growth. These determinants of economic growth can be grouped into four supply factors, one demand factor, and one efficiency factor.

Supply Factors

The first four determinants of economic growth relate to the physical ability of the economy to expand. They are:

- Increases in the quantity and quality of natural resources.
- Increases in the quantity and quality of human resources.
- Increases in the supply (or stock) of capital goods.
- Improvements in technology.

Any increases or improvements in these **supply factors** will increase the *potential* size of an economy's GDP. The remaining two factors are necessary for that potential to be fulfilled not just in terms of the overall quantity of output but also in terms of the quality of that output and whether it is properly directed toward producing the items most highly valued by society.

Demand Factor

The fifth determinant of economic growth is the **demand factor:**

- To actually achieve the higher production potential created when the supply factors increase or improve, households, businesses, and the government must also expand their purchases of goods and services so as to provide a market for all the new output that can potentially be produced.

If that occurs, there will be no unplanned increases in inventories and resources will remain fully employed. The demand factor acknowledges that economic growth requires that increases in total spending must occur if we are to actually realize the output gains made available by increased production capacity.

Efficiency Factor

The sixth determinant of economic growth is the **efficiency factor:**

- To reach its full production potential, an economy must achieve economic efficiency as well as full employment.

The economy must use its resources in the least costly way (productive efficiency) to produce the specific mix of goods and services that maximizes people's well-being (allocative efficiency). The ability to expand production, together with the full use of available resources, is not sufficient for achieving maximum possible growth. Also required is the efficient use of those resources.

The supply, demand, and efficiency factors in economic growth are related. Unemployment caused by insufficient total spending (the demand factor) may lower the rate of new capital accumulation (a supply factor) and delay expenditures on research (also a supply factor). Conversely, low spending on investment (a supply factor) may cause insufficient spending (the demand factor) and unemployment. Widespread inefficiency in the use of resources (the efficiency factor) may translate into higher costs of goods and services and thus lower profits, which in turn may slow innovation and reduce the accumulation of capital (supply factors). Economic growth is a dynamic process in which the supply, demand, and efficiency factors all interact.

ORIGIN OF THE IDEA

O26.1

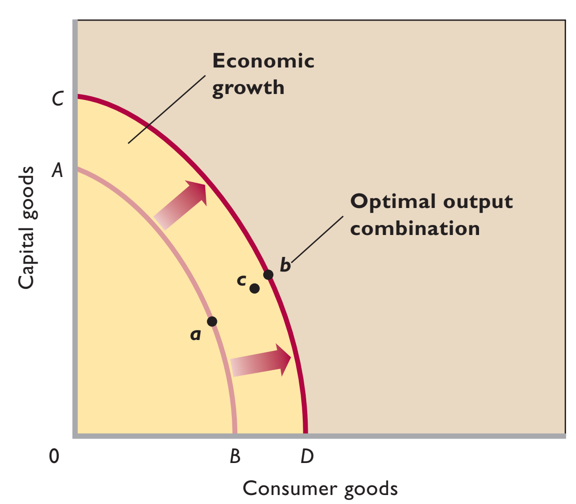
Growth theory

Production Possibilities Analysis

To put the six factors affecting the rate of economic growth into better perspective, let's use the production possibilities analysis introduced in Chapter 1.

Growth and Production Possibilities Recall that a curve like *AB* in Figure 26.2 is a production possibilities

FIGURE 26.2 Economic growth and the production possibilities curve. Economic growth is made possible by the four supply factors that shift the production possibilities curve outward, as from *AB* to *CD*. Economic growth is realized when the demand factor and the efficiency factor move the economy from points such as *a* and *c* that are inside *CD* to the optimal output point, which is assumed to be point *b* in this figure.



curve. It indicates the various *maximum* combinations of products an economy can produce with its fixed quantity and quality of natural, human, and capital resources and its stock of technological knowledge. An improvement in any of the supply factors will push the production possibilities curve outward, as from *AB* to *CD*.

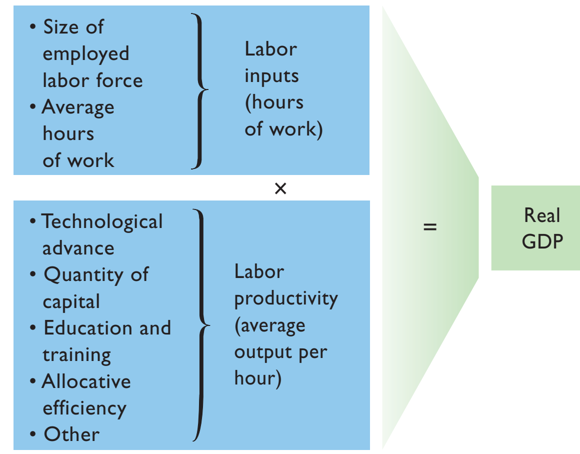
But the demand factor reminds us that an increase in total spending is needed to move the economy from a point like *a* on curve *AB* to any of the points on the higher curve *CD*. And the efficiency factor reminds us that we need least-cost production and an optimal location on *CD* for the resources to make their maximum possible dollar contribution to total output. You will recall from Chapter 1 that this “best allocation” is determined by expanding the production of each good until its marginal benefit equals its marginal cost. Here, we assume that this optimal combination of capital and consumer goods occurs at point *b*. If the efficiency factor is in full effect, then the economy will produce at point *b* rather than at any other point along curve *CD*.

Example: The net increase in the size of the labor force in the United States in recent years has been 1.5 to 2 million workers per year. That increment raises the economy’s production capacity. But obtaining the extra output that these added workers could produce depends on their success in finding jobs. It also depends on whether or not the jobs are in firms and industries where the workers’ talents are fully and optimally used. Society does not want new labor-force entrants to be unemployed. Nor does it want pediatricians working as plumbers or pediatricians producing pediatric services for which marginal costs exceed marginal benefits.

Normally, increases in total spending match increases in production capacity, and the economy moves from a point on the previous production possibilities curve to a point on the expanded curve. Moreover, the competitive market system tends to drive the economy toward productive and allocative efficiency. Occasionally, however, the economy may end up at some point such as *c* in Figure 26.2. That kind of outcome occurred in the United States during the severe recession of 2007–2009. Real output fell far below the amount of output that the economy could have produced if it had achieved full employment and operated on its production possibilities curve.

Labor and Productivity Although the demand and efficiency factors are important, discussions of economic growth focus primarily on supply factors. Society can increase its real output and income in two fundamental ways: (1) by increasing its inputs of resources and (2) by raising the productivity of those inputs. Figure 26.3 concentrates

FIGURE 26.3 The supply determinants of real output. Real GDP is usefully viewed as the product of the quantity of labor inputs (hours of work) multiplied by labor productivity.



on the input of *labor* and provides a useful framework for discussing the role of supply factors in growth. A nation’s real GDP in any year depends on the input of labor (measured in hours of work) multiplied by **labor productivity** (measured as real output per hour of work):

$$\text{Real GDP} = \text{hours of work} \times \text{labor productivity}$$

Thought of this way, a nation’s economic growth from one year to the next depends on its *increase* in labor inputs (if any) and its *increase* in labor productivity (if any).

Illustration: Assume that the hypothetical economy of Ziam has 10 workers in year 1, each working 2,000 hours per year (50 weeks at 40 hours per week). The total input of labor therefore is 20,000 hours. If productivity (average real output per hour of work) is \$10, then real GDP in Ziam will be \$200,000 (= 20,000 × \$10). If work hours rise to 20,200 and labor productivity rises to \$10.40, Ziam’s real GDP will increase to \$210,080 in year 2. Ziam’s rate of economic growth will be about 5 percent [= (\$210,080 − \$200,000)/\$200,000] for the year.

WORKED PROBLEMS

W26.2
Productivity and economic growth

Hours of Work What determines the number of hours worked each year? As shown in Figure 26.3, the hours of labor input depend on the size of the employed labor force and the length of the average workweek. Labor-force size depends on the size of the working-age population and the **labor-force participation rate**—the percentage of the working-age population actually in the labor force. The length of the average workweek is governed by legal and institutional considerations and by collective bargaining agreements negotiated between unions and employers.

Labor Productivity Figure 26.3 tells us that labor productivity is determined by technological progress, the quantity of capital goods available to workers, the quality of labor itself, and the efficiency with which inputs are allocated, combined, and managed. Productivity rises when the health, training, education, and motivation of workers improve; when workers have more and better machinery and natural resources with which to work; when production is better organized and managed; and when labor is reallocated from less-efficient industries to more-efficient industries.

Accounting for Growth

LO26.4 Describe "growth accounting" and the specific factors accounting for economic growth in the United States.

The president's Council of Economic Advisers uses a system called **growth accounting** to assess the relative importance of the supply-side elements that contribute to changes in real GDP. This system groups these elements into two main categories:

- Increases in hours of work.
- Increases in labor productivity.

Labor Inputs versus Labor Productivity

Table 26.3 provides the relevant data for the United States for five periods. The symbol "Q" in the table stands for "quarter" of the year. The beginning points for the first four periods are business-cycle peaks, and the last period includes future projections by the Council of Economic Advisers. It is clear from the table that both increases in the quantity of labor and increases in labor productivity are important sources of economic growth. Between 1953 and 2012, the labor force increased from 63 million to 155 million workers. Over that period the average length of the workweek remained relatively stable. Falling birthrates slowed the growth of the native population, but increased immigration partly offset that slowdown. As indicated in the Consider This box on the next page, of particular significance was a surge of women's participation in the labor force. Partly as a result, U.S. labor-force growth averaged 1.6 million workers per year over those 56 years.

The growth of labor productivity also has been important to economic growth. In fact, productivity growth has usually been the more significant factor, with the exception of 1973–1995 when productivity growth greatly slowed. For example, between 2001 and 2011, productivity growth was responsible for all of the 1.7 percent average annual economic rate because labor inputs were shrinking over that time period. The size of the labor force did increase over that decade, but fewer total hours were worked due to many workers shifting from full-time to part-time work and because of the high rates of unemployment experienced during and after that decade's two recessions (in 2001 and 2007–2009). As shown in the far right column of Table 26.3, productivity growth is projected to account for 92 percent of the growth of real GDP between 2011 and 2021.

Because increases in labor productivity are so important to economic growth, economists go to the trouble of investigating and assessing the relative importance of the factors that contribute to productivity growth. There are five factors that, together, appear to explain changes in productivity growth rates: technological advance, the amount of capital each worker has to work with, education and training, economies of scale, and resource allocation. We will examine each factor in turn, noting how much each factor contributes to productivity growth.

Technological Advance

The largest contributor to productivity growth is technological advance, which is thought to account for about

TABLE 26.3 **Accounting for the Growth of U.S. Real GDP, 1953–2011 Plus Projection from 2011 to 2022 (Average Annual Percentage Changes)**

	Actual				Projected
Item	1953 Q2 to 1973 Q4	1973 Q4 to 1995 Q2	1995 Q2 to 2001 Q1	2001 Q1 to 2011 Q1	2011 Q1 to 2021 Q4
Increase in real GDP	3.6	2.8	3.8	1.7	2.5
Increase in quantity of labor	1.1	1.3	1.4	−0.7	0.2
Increase in labor productivity	2.5	1.5	2.4	2.4	2.3

Source: Derived from *Economic Report of the President, 2008,* p. 45; *Economic Report of the President, 2010,* p. 76; *Economic Report of the President 2011,* p. 52; Bureau of Economic Analysis; and Bureau of Labor Statistics.

40 percent of productivity growth. As economist Paul Romer stated, "Human history teaches us that economic growth springs from better recipes, not just from more cooking."

Technological advance includes not only innovative production techniques but new managerial methods and new forms of business organization that improve the process of production. Generally, technological advance is generated by the discovery of new knowledge, which allows resources to be combined in improved ways that increase output. Once discovered and implemented, new knowledge soon becomes available to entrepreneurs and firms at relatively low cost. Technological advance therefore eventually spreads through the entire economy, boosting productivity and economic growth.

Technological advance and capital formation (investment) are closely related, since technological advance usually promotes investment in new machinery and equipment. In fact, technological advance is often *embodied* within new capital. For example, the purchase of new computers brings into industry speedier, more powerful computers that incorporate new technology.

Technological advance has been both rapid and profound. Gas and diesel engines, conveyor belts, and assembly lines are significant developments of the past. So, too, are fuel-efficient commercial aircraft, integrated microcircuits, personal computers, digital photography, and containerized shipping. More recently, technological advance has exploded, particularly in the areas of computers, photography, wireless communications, and the Internet. Other fertile areas of recent innovation are medicine and biotechnology.

Quantity of Capital

A second major contributor to productivity growth is increased capital, which explains roughly 30 percent of productivity growth. More and better plant and equipment make workers more productive. And a nation acquires more capital by saving some of its income and using that savings to invest in plant and equipment.

Although some capital substitutes for labor, most capital is complementary to labor—it makes labor more productive. A key determinant of labor productivity is the amount of capital goods available *per worker*. If both the aggregate stock of capital goods and the size of the labor force increase over a given period, the individual worker is not necessarily better equipped and productivity will not necessarily rise. But the quantity of capital equipment available per U.S. worker has increased greatly over time. (In 2011 it was about $126,062 per worker.)

CONSIDER THIS . . .

Women, the Labor Force, and Economic Growth

The substantial rise in the number of women working in the paid workforce in the United States has been one of the major labor market trends of the last 50 years. In 1960, about 40 percent of women worked full-time or part-time in paid jobs. Today, that number is about 60 percent.

Women have greatly increased their productivity in the workplace, mostly by becoming better educated and professionally trained. Rising productivity has increased women's wage rates. Those higher wages have raised the opportunity costs—the forgone wage earnings—of staying at home. Women have therefore substituted employment in the labor market for traditional home activities. This substitution has been particularly pronounced among married women. (Single women have always had high labor-force participation rates.)

Furthermore, changing lifestyles and the widespread availability of birth control have freed up time for greater labor-force participation by women. Women not only have fewer children, but those children are spaced closer together in age. Thus women who leave their jobs during their children's early years return to the labor force sooner.

Greater access to jobs by women also has raised the labor-force participation of women. Service industries—teaching, nursing, and office work, for instance—that traditionally have employed many women have expanded rapidly in the past several decades. Also, the population in general has shifted from farms and rural regions to urban areas, where jobs for women are more abundant and more geographically accessible. Additionally, occupational barriers to professions have greatly eroded, resulting in many more women becoming business managers, lawyers, professors, and physicians.

In summary, women in the United States are better educated, more productive, and more efficiently employed than ever before. Their greater presence in the labor force has contributed greatly to U.S. economic growth.

Public investment in the U.S. **infrastructure** (highways and bridges, public transit systems, wastewater treatment facilities, water systems, airports, educational facilities, and so on) has also grown over the years. This publicly owned capital complements private capital. Investments in new highways promote private investment in new factories and retail stores along their routes.

Industrial parks developed by local governments attract manufacturing and distribution firms.

Private investment in infrastructure also plays a large role in economic growth. One example is the tremendous growth of private capital relating to communications systems over the years.

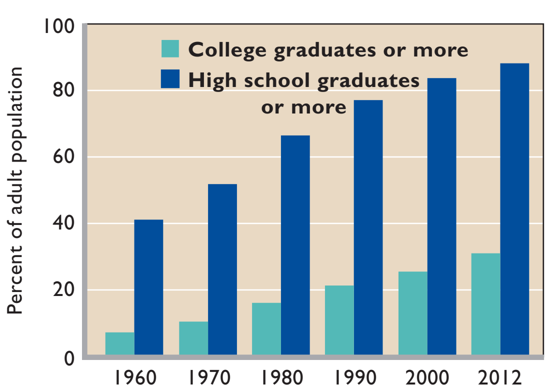
Education and Training

Ben Franklin once said, "He that hath a trade hath an estate," meaning that education and training contribute to a worker's stock of **human capital**—the knowledge and skills that make a worker productive. Investment in human capital includes not only formal education but also on-the-job training. Like investment in physical capital, investment in human capital is an important means of increasing labor productivity and earnings. An estimated 15 percent of productivity growth derives from investments in people's education and skills.

One measure of a nation's quality of labor is its level of educational attainment. Figure 26.4 shows large gains in education attainment over the past several decades. In 1960 only 41 percent of the U.S. population age 25 or older had at least a high school education; and only 8 percent had a college or postcollege education. By 2012, those numbers had increased to 88 and 31 percent, respectively. Clearly, more people are receiving more education than ever before.

But all is not upbeat with education in the United States. Many observers think that the quality of education in the United States has declined. For example, U.S. students perform poorly on science and math tests relative to students in many other nations (see Global Perspective 26.1). And the United States has been producing fewer engineers and scientists, a problem that may trace back to inadequate training in math and science in elementary and high schools. For these reasons, much recent public policy discussion and legislation have been directed toward improving the quality of the U.S. education and training system.

FIGURE 26.4 Changes in the educational attainment of the U.S. adult population. The percentage of the U.S. adult population, age 25 or older, completing high school and college has been rising over recent decades.



Source: U.S. Census Bureau, **www.census.gov**.

GLOBAL PERSPECTIVE 26.1

Average Test Scores of Eighth-Grade Students in Math and Science, Top 10 Test-Taking Countries

The test performance of U.S. eighth-grade students did not compare favorably with that of eighth-graders in several other nations in the Fifth International Math and Science Study (2011).

Mathematics

Rank		Score
1	South Korea	613
2	Singapore	611
3	Taiwan	609
4	Hong Kong	586
5	Japan	570
6	Russia	539
7	Israel	516
8	Finland	514
9	United States	509
10	Great Britain	507

Science

Rank		Score
1	Singapore	590
2	Taiwan	564
3	South Korea	560
4	Japan	558
5	Finland	552
6	Slovenia	543
7	Russia	542
8	Hong Kong	535
9	Great Britain	533
10	United States	525

Economies of Scale and Resource Allocation

Economies of scale and improved resource allocation are a fourth and fifth source of productivity growth, and together they explain about 15 percent of productivity growth.

Economies of Scale Reductions in per-unit production costs that result from increases in output levels are called **economies of scale.** Markets have increased in size over time, allowing firms to increase output levels and thereby achieve production advantages associated with greater size. As firms expand their size and output, they are able to use larger, more productive equipment and employ methods of manufacturing and delivery that increase productivity. They also are better able to recoup substantial investments in developing new products and production methods. Examples: A large manufacturer of autos can use elaborate assembly lines with computerization and robotics, while smaller producers must settle for less-advanced technologies using more labor inputs. Large pharmaceutical firms greatly reduce the average amount of labor (researchers, production workers) needed to produce each pill as they increase the number of pills produced. Accordingly, economies of scale result in greater real GDP and thus contribute to economic growth.

Improved Resource Allocation Improved resource allocation means that workers over time have moved from low-productivity employment to high-productivity employment. Historically, many workers have shifted from agriculture, where labor productivity is low, to manufacturing, where it is quite high. More recently, labor has shifted away from some manufacturing industries to even higher-productivity industries such as computer software, business consulting, and pharmaceuticals. As a result of such shifts, the average productivity of U.S. workers has increased.

Also, discrimination in education and the labor market has historically deterred some women and minorities from entering high-productivity jobs. With the decline of such discrimination over time, many members of those groups have shifted from lower-productivity jobs to higher-productivity jobs. The result has been higher overall labor productivity and real GDP.

Finally, things such as tariffs, import quotas, and other barriers to international trade tend to relegate resources to relatively unproductive pursuits. The long-run movement toward liberalized international trade through international agreements has improved the allocation of resources, increased labor productivity, and expanded real output, both here and abroad.

QUICK REVIEW 26.3

- Institutional structures that promote growth include strong property rights, patents, efficient financial institutions, education, and a competitive market system.
- The determinants of economic growth include four supply factors (increases in the quantity and quality of natural resources, increases in the quantity and quality of human resources, increases in the stock of capital goods, and improvements in technology); one demand factor (increases in total spending); and one efficiency factor (achieving allocative and productive efficiency).
- Improvements in labor productivity accounted for about two-thirds of the increase in U.S. real GDP between 1990 and 2012; the use of more labor inputs accounted for the remainder.
- Improved technology, more capital, greater education and training, economies of scale, and better resource allocation have been the main contributors to U.S. productivity growth and thus to U.S. economic growth.

The Rise in the Average Rate of Productivity Growth

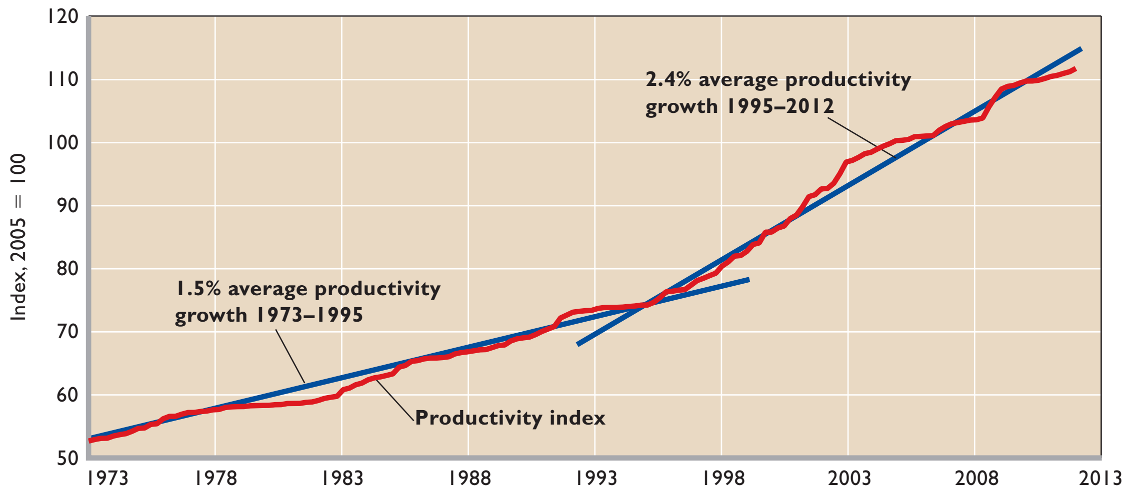
LO26.5 Explain why the trend rate of U.S. productivity growth has increased since the earlier 1973–1995 period.

Figure 26.5 shows the growth of labor productivity (as measured by changes in the index of labor productivity) in the United States from 1973 to 2012, along with separate trend lines for 1973–1995 and 1995–2012. Labor productivity in the business sector grew by an average of only 1.5 percent yearly over the 1973–1995 period. But productivity growth averaged 2.4 percent between 1995 and 2012. Many economists believe that this higher productivity growth resulted from a significant new wave of technological advance, coupled with global competition. Some economists think there is a good chance that the higher trend rates of productivity growth could continue for many years to come.

This increase in productivity growth is important because real output, real income, and real wages are linked to labor productivity. To see why, suppose you are alone on an uninhabited island. The number of fish you can catch or coconuts you can pick per hour—your productivity—is your real wage (or real income) per hour. By *increasing* your productivity, you can improve your standard of living because you can gather more fish and more coconuts (goods) for each hour of work.

So it is for the economy as a whole: Over long periods, the economy's labor productivity determines its average real hourly wage, which includes fringe benefits such as health

FIGURE 26.5 Growth of labor productivity in the United States, 1973–2012. U.S. labor productivity (here, for the business sector) increased at an average annual rate of only 1.5 percent from 1973 to 1995. But between 1995 and 2012, it rose at an annual rate of 2.4 percent.



Source: U.S. Bureau of Labor Statistics, **www.bls.gov**.

care insurance and contributions to pensions. The economy's income per hour is equal to its output per hour. So productivity growth is the economy's main route for improving the living standards for its workers. It allows firms to pay higher wages without lowering their business profits.

Reasons for the Rise in the Average Rate of Productivity Growth

Why has productivity growth increased relative to earlier periods?

The Microchip and Information Technology The core element of the productivity speedup is an explosion of entrepreneurship and innovation based on the microprocessor, or *microchip*, which bundles transistors on a piece of silicon. Some observers liken the invention of the microchip to that of electricity, the automobile, air travel, the telephone, and television in importance and scope.

The microchip has found its way into thousands of applications. It has helped create a wide array of new products and services and new ways of doing business. Its immediate results were the pocket calculator, the bar-code scanner, the personal computer, the laptop computer, and more powerful business computers. But the miniaturization of electronic circuits also advanced the development of many other products such as cell phones and pagers, computer-guided lasers, global positioning equipment, energy conservation systems, Doppler radar, digital cameras, and machines to decipher the human genome.

Perhaps of greatest significance, the widespread availability of personal and laptop computers stimulated the desire to tie them together. That desire promoted rapid development of the Internet and all its many manifestations, such as business-to-household and business-to-business electronic commerce (e-commerce). The combination of the computer, fiber-optic cable, wireless technology, and the Internet constitutes a spectacular advance in **information technology,** which has been used to connect all parts of the world.

New Firms and Increasing Returns Hundreds of new **start-up firms** advanced various aspects of the new information technology. Many of these firms created more "hype" than goods and services and quickly fell by the wayside. But a number of firms flourished, eventually to take their places among the nation's largest firms. Examples of those firms include Intel (microchips); Apple and Dell (personal computers); Microsoft and Oracle (computer software); Cisco Systems (Internet switching systems); America Online (Internet service provision); Yahoo and Google (Internet search engines); and eBay, PayPal, and Amazon.com (electronic commerce). There are scores more! Most of these firms were either "not on the radar" or "a small blip on the radar" 30 years ago. Today each of them has large annual revenue and employs thousands of workers.

Successful new firms often experience **increasing returns,** a situation in which a given percentage increase in the amount of inputs a firm uses leads to an even larger percentage increase in the amount of output the firm produces. For example, suppose that a company called Techco decides to double the size of its operations to meet the growing demand for its services. After doubling its plant and equipment and doubling its workforce, say, from

100 workers to 200 workers, it finds that its total output has tripled from 8,000 units to 24,000 units. Techco has experienced increasing returns; its output has increased by 200 percent, while its inputs have increased by only 100 percent. That is, its labor productivity has gone up from 80 units per worker (= 8,000 units/100 workers) to 120 units per worker (= 24,000 units/200 workers). Increasing returns boost labor productivity and reduce per-unit production costs. Since these cost reductions result from increases in output levels, they are examples of *economies of scale*.

Both emerging firms as well as established firms can exploit several different sources of increasing returns and economies of scale:

- ***More specialized inputs*** Firms can use more specialized and thus more productive capital and workers as they expand their operations. A growing new e-commerce business, for example, can purchase highly specialized inventory management systems and hire specialized personnel such as accountants, marketing managers, and system maintenance experts.
- ***Spreading of development costs*** Firms can spread high product development costs over greater output. For example, suppose that a new software product costs $100,000 to develop and only $2 per unit to manufacture and sell. If the firm sells 1,000 units of the software, its per-unit cost will be $102 [= ($100,000 + $2,000)/1,000], but if it sells 500,000 units, that cost will drop to only $2.20 [= ($100,000 + $1 million)/500,000].
- ***Simultaneous consumption*** Many recently developed products and services can satisfy large numbers of customers at the same time. Unlike a gallon of gas that needs to be produced for each buyer, a software program needs to be produced only once. It then becomes available at very low expense to thousands or even millions of buyers. The same is true of books delivered to electronic reading devices, movies distributed on DVDs, and information disseminated through the Internet.
- ***Network effects*** Software and Internet service become more beneficial to a buyer the greater the number of households and businesses that also buy them. When others have Internet service, you can send e-mail messages to them. And when they also have software that allows display of documents and photos, you can attach those items to your e-mail messages. These interconnectivity advantages are called **network effects,** which are increases in the value of a product to each user, including existing users, as the total number of users rises. The domestic and global expansion of the Internet in particular has produced network effects, as have cell phones, pagers, tablet computers, and other aspects of wireless communication. Network effects magnify the value of output well beyond the costs of inputs.
- ***Learning by doing*** Finally, firms that produce new products or pioneer new ways of doing business experience increasing returns through **learning by doing.** Tasks that initially may have taken firms hours may take them only minutes once the methods are perfected.

Whatever the particular source of increasing returns, the result is higher productivity, which tends to reduce the per-unit cost of producing and delivering products.

Global Competition The recent economy is characterized not only by information technology and increasing returns but also by heightened global competition. The collapse of the socialist economies in the late 1980s and early 1990s, together with the success of market systems, has led to a reawakening of capitalism throughout the world. The new information technologies have "shrunk the globe" and made it imperative for all firms to lower their costs and prices and to innovate in order to remain competitive. Free-trade zones such as NAFTA and the European Union (EU), along with trade liberalization through the World Trade Organization (WTO), have also heightened competition internationally by removing trade protection from domestic firms. The larger geographic markets, in turn, have enabled firms to expand beyond their national borders.

Implications for Economic Growth

Other things equal, stronger productivity growth and heightened global competition allow the economy to achieve a higher rate of economic growth. A glance back at Figure 26.2 will help make this point. Suppose that the shift of the production possibilities curve from *AB* to *CD* reflects annual changes in potential output levels before the recent increase in growth rates. Then the higher growth rates of the more recent period of accelerated productivity growth would be depicted by a *larger* outward shift of the economy's production possibilities from *AB* to a curve beyond *CD*. When coupled with economic efficiency and increased total spending, the economy's real GDP would rise by even more than what is shown.

Two cautions: Although the trend line of productivity growth seems to be steeper than in the past and bodes well for long-term economic growth, fluctuations of the rate of economic growth will still occur. Because of demand factors, real output periodically deviates below and above the growth trend—as it certainly did during the recession of 2007–2009. Also, you need to know that the growth of the U.S. labor force may be declining. That slowing may offset some or all of the extra potential for economic growth that would arise from greater productivity growth.

Skepticism about Longevity

Although most macroeconomists have revised their forecasts for long-term productivity growth upward, at least slightly, others are still skeptical and urge a "wait-and-see" approach. These macroeconomists acknowledge that the economy has experienced a rapid advance of new technology, some new firms have experienced increasing returns, and global competition has increased. But they wonder if these factors are sufficiently profound to produce a long-lasting new era of substantially higher rates of productivity growth and real GDP growth.

They also point out that productivity surged between 1975 and 1978 and between 1983 and 1986 but in each case soon reverted to its lower long-run trend. The higher trend line of productivity inferred from the short-run spurt of productivity could prove to be transient. Only by looking backward over long periods can economists distinguish the start of a new long-run trend from a shorter-term boost in productivity related to the business cycle and temporary factors.

What Can We Conclude?

Given the different views on the recent productivity acceleration, what should we conclude? Perhaps the safest conclusions are these:

- The prospects for a lasting increase in productivity growth are good (see Global Perspective 26.2). Studies indicate that productivity increases related to information technology have spread to a wide range of industries, including services. Even during the severe 2007–2009 recession, when real GDP fell by nearly 5 percent, productivity growth continued. Specifically, it ran at a 2.1 percent rate in 2007 just as the recession was starting, fell to 0.9 percent in 2008 when the recession was at its worst, and then jumped up to a very brisk 4.9 percent as the economy began to recover in 2009.
- Time will tell. After the rapid productivity growth of 2009, the growth rate fell to just 0.5 percent in 2010, 0.9 percent in 2011, and 1.0 percent in 2012. Consequently, the average growth rate over those four years is lower than the 2.4 percent rate shown by the trend line in Figure 26.5—but it is higher than the 1.5 percent rate for the 1973–1995 period. Clearly, many more years must elapse before economists will be ready to declare the post-1995 productivity acceleration a long-run, sustainable trend.

GLOBAL PERSPECTIVE 26.2

Global Competitiveness Index

The Global Competitiveness Index, published annually by the World Economic Forum, measures each country's potential for economic growth. The index uses various factors—such as innovativeness, the capability to transfer technology among sectors, the efficiency of the financial system, rates of investment, and the degree of integration with the rest of the world—to measure a country's ability to achieve economic growth over time. Here is the top 10 list for 2012–2013.

Country	Global Competitiveness Ranking, 2012–2013
Switzerland	1
Singapore	2
Finland	3
Sweden	4
Netherlands	5
Germany	6
United States	7
United Kingdom	8
Hong Kong	9
Japan	10

Source: Copyright World Economic Forum, **www.weforum.org**.

QUICK REVIEW 26.4

- Over long time periods, labor productivity growth determines an economy's growth of real wages and its standard of living.
- Many economists believe that the United States has entered a period of faster productivity growth and possibly higher rates of economic growth.
- The rise in the average rate of productivity growth is based on rapid technological change in the form of the microchip and information technology, increasing returns and lower per-unit costs, and heightened global competition that helps hold down prices.
- More-rapid U.S. productivity growth means that, other things equal, the U.S. economy can grow at higher annual rates than it could with less-rapid productivity growth. Nonetheless, many economists caution that it is still too early to determine whether the higher rates of productivity growth since 1995 are a lasting long-run trend or a fortunate short-lived occurrence.

Is Growth Desirable and Sustainable?

LO26.6 Discuss differing perspectives as to whether growth is desirable and sustainable.

Economists usually take for granted that economic growth is desirable and sustainable. But not everyone agrees.

The Antigrowth View

Critics of growth say industrialization and growth result in pollution, climate change, ozone depletion, and other environmental problems. These adverse negative externalities occur because inputs in the production process reenter the environment as some form of waste. The more rapid our growth and the higher our standard of living, the more waste the environment must absorb—or attempt to absorb. In an already wealthy society, further growth usually means satisfying increasingly trivial wants at the cost of mounting threats to the ecological system.

Critics of growth also argue that there is little compelling evidence that economic growth has solved sociological problems such as poverty, homelessness, and discrimination. Consider poverty: In the antigrowth view, American poverty is a problem of distribution, not production. The requisite for solving the problem is a firm commitment to redistribute wealth and income, not further increases in output.

Antigrowth sentiment also says that while growth may permit us to "make a better living," it does not give us "the good life." We may be producing more and enjoying it less. Growth means frantic paces on jobs, worker burnout, and alienated employees who have little or no control over decisions affecting their lives. The changing technology at the core of growth poses new anxieties and new sources of insecurity for workers. Both high-level and low-level workers face the prospect of having their hard-earned skills and experience rendered obsolete by onrushing technology. High-growth economies are high-stress economies, which may impair our physical and mental health.

Finally, critics of high rates of growth doubt that they are sustainable. The planet Earth has finite amounts of natural resources available, and they are being consumed at alarming rates. Higher rates of economic growth simply speed up the degradation and exhaustion of the earth's resources. In this view, slower economic growth that is environmentally sustainable is preferable to faster growth.

In Defense of Economic Growth

The primary defense of growth is that it is the path to the greater material abundance and higher living standards desired by the vast majority of people. Rising output and incomes allow people to buy

> more education, recreation, and travel, more medical care, closer communications, more skilled personal and professional services, and better-designed as well as more numerous products. It also means more art, music, and poetry, theater, and drama. It can even mean more time and resources devoted to spiritual growth and human development.[1]

Growth also enables society to improve the nation's infrastructure, enhance the care of the sick and elderly, provide greater access for the disabled, and provide more police and fire protection. Economic growth may be the only realistic way to reduce poverty, since there is only limited political support for greater redistribution of income. The way to improve the economic position of the poor is to increase household incomes through higher productivity and economic growth. Also, a no-growth policy among industrial nations might severely limit growth in poor nations. Foreign investment and development assistance in those nations would fall, keeping the world's poor in poverty longer.

Economic growth has not made labor more unpleasant or hazardous, as critics suggest. New machinery is usually less taxing and less dangerous than the machinery it replaces. Air-conditioned workplaces are more pleasant than steamy workshops. Furthermore, why would an end to economic growth reduce materialism or alienation? The loudest protests against materialism are heard in those nations and groups that now enjoy the highest levels of material abundance! The high standard of living that growth provides has increased our leisure and given us more time for reflection and self-fulfillment.

Does growth threaten the environment? The connection between growth and environment is tenuous, say growth proponents. Increases in economic growth need not mean increases in pollution. Pollution is not so much a by-product of growth as it is a "problem of the commons." Much of the environment—streams, lakes, oceans, and the air—is treated as common property, with insufficient or no restrictions on its use. The commons have become our dumping grounds; we have overused

[1] Alice M. Rivlin, *Reviving the American Dream* (Washington, D.C.: Brookings Institution, 1992), p. 36.

LAST WORD

Can Economic Growth Survive Population Decline?

The Demographic Transition Is Causing Greying Populations, Shrinking Labor Forces, and Overall Population Decreases in Many Nations. Can Economic Growth Survive?

As you know from this chapter, Real GDP = hours of work × labor productivity. The number of *hours of work* depends heavily, however, on the size of the working-age population. If it begins to shrink, the number of *hours of work* almost always falls. In such cases, the only way real GDP can rise is if *labor productivity* increases faster than *hours of work* decreases. The world is about to see if that can happen in countries that have populations that are greying and shrinking.

The historical background has to do with the fact that as nations industrialize, their economies shift from agriculture to industry. As that happens, fertility levels plummet because the shift to modern technology transforms children from being economically essential farm hands that can contribute to their families' incomes from a young age to expensive investment goods that require many years of costly schooling before they can support themselves.

As people react to this change, birthrates tend to fall quite dramatically. The key statistic is the *total fertility rate* that keeps track of the average number of births that women have during their lifetimes. To keep the population stable in modern societies, the total fertility rate must be about 2.1 births per woman per lifetime (= 1 child to replace mom, 1 child to replace dad, and 0.1 child to compensate for those people who never end up reproducing as adults).

Every rich industrial nation has now seen its total fertility rate drop below the replacement level of 2.1 births per woman per lifetime. In Japan and many Eastern European countries, the number has been so low for so long that

and debased them. Environmental pollution is a case of negative externalities, and correcting this problem involves regulatory legislation, specific taxes ("effluent charges"), or market-based incentives to remedy misuse of the environment.

Those who support growth admit there are serious environmental problems. But they say that limiting growth is the wrong solution. Growth has allowed economies to reduce pollution, be more sensitive to environmental considerations, set aside wilderness, create national parks and monuments, and clean up hazardous waste, while still enabling rising household incomes. (See the Last Word in Chapter 17.)

Is growth sustainable? Yes, say the proponents of growth. If we were depleting natural resources faster than their discovery, we would see the prices of those resources rise. That has not been the case for most natural resources; in fact, the prices of most of them have declined (see Figure 17.1). And if one natural resource becomes too expensive, another resource will be substituted for it. Moreover, say economists, economic growth has to do with the expansion and application of human knowledge and information, not of extractable natural resources. In this view, economic growth is limited only by human imagination.

QUICK REVIEW 26.5

- Critics of growth argue that it adds to environmental degradation, increases human stress, and exhausts the earth's finite supply of natural resources.
- Defenders of growth say that it is the primary path to the rising living standards, that it need not debase the environment, and that there are no indications that we are running out of resources.
- Defenders of growth argue that it is sustainable because growth is based on the expansion and application of human knowledge, which is limited only by human imagination.

there are no longer enough children being born each year to replace the old folks who are dying. As a result, their overall populations are shrinking.

Economists only expect that pattern to become more common and more rapid, so that by the year 2050 the majority of nations will have decreasing populations. But decades before a nation's overall population begins to decrease, it faces a situation in which the labor force shrinks while the elderly population swells.

That pattern is the result of each generation being smaller than the one before. As an example, the Baby Boom generation born between 1946 and 1964 is much larger than the Baby Bust generation that followed it. So as the Boomers retire over the next two decades, there will be a lot of retirees as compared to working-age adults.

This trend can be quantified by the *inverse dependency ratio*, which is defined as the number of people of working age (ages 20 to 64) divided by the number of dependents (seniors over age 65 plus youths under age 20). In the United States, the inverse dependency ratio is set to fall from 1.5 people of working age per dependent in 2010 to just 1.16 people of working age per dependent in 2050. That is extremely problematic because it implies that worker productivity will have to rise dramatically just to make up for the relative decline in the number of workers as compared to dependents. If productivity doesn't keep up with the fall in the inverse dependency ratio, living standards will have to decline because there will simply be too many nonworking consumers relative to working-age producers.

The place where this problem is likely to show up first is Social Security. There are currently 2.9 workers paying into the Social Security system for each retiree receiving Social Security benefits. But that number is set to fall to just 2.0 workers per retiree in 2030. So worker productivity would have to increase by almost a third in under 20 years just to keep up with the decline in the number of workers relative to retirees.

Economists are uncertain about whether such large productivity increases will be forthcoming. The problem is that consumption competes with investment. A society with a larger fraction of dependents is a society that is likely to devote an increasingly high fraction of total output toward consumption rather than investment. If so, productivity growth may slow considerably.

Another possible problem is that, historically, most transformative new technologies and businesses have been created by energetic young people under the age of 40. With each generation getting smaller, there will be fewer people in that age range and thus, possibly, less innovation and slower productivity growth.

Other economists are more hopeful, however. They view old people as consumers and demanders. As their numbers swell, inventors may simply switch from inventing products for young people to inventing products for old people. If so, productivity growth and living standards could keep on rising at the rates we have come to expect.

SUMMARY

LO26.1 List two ways that economic growth is measured.

A nation's economic growth can be measured either as an increase in real GDP over time or as an increase in real GDP per capita over time. Real GDP in the United States has grown at an average annual rate of about 3.2 percent since 1950; real GDP per capita has grown at roughly a 2 percent annual rate over that same period.

LO26.2 Define "modern economic growth" and explain the institutional structures needed for an economy to experience it.

Sustained increases in real GDP per capita did not happen until the past two centuries, when England and then other countries began to experience modern economic growth, which is characterized by institutional structures that encourage savings, investment, and the development of new technologies. Institutional structures that promote growth include strong property rights, patents, efficient financial institutions, education, and a competitive market system.

Because some nations have experienced nearly two centuries of modern economic growth while others have only recently begun to experience modern economic growth, some countries today are much richer than other countries.

It is possible, however, for countries that are currently poor to grow faster than countries that are currently rich because the growth of real GDP per capita for rich countries is limited to about 2 percent per year. To continue growing, rich countries must invent and apply new technologies. By contrast, poor countries can grow much faster because they can simply adopt the institutions and cutting-edge technologies already developed by the rich countries.

LO26.3 Identify the general supply, demand, and efficiency forces that give rise to economic growth.

The determinants of economic growth to which we can attribute changes in growth rates include four supply factors (changes in the quantity and quality of natural resources, changes in the quantity and quality of human resources, changes in the stock of capital goods, and improvements in technology); one demand factor (changes in total spending); and one efficiency factor (changes in how well an economy achieves allocative and productive efficiency).

The growth of a nation's capacity to produce output can be illustrated graphically by an outward shift of its production possibilities curve.

LO26.4 Describe "growth accounting" and the specific factors accounting for economic growth in the United States.

Growth accounting attributes increases in real GDP either to increases in the amount of labor being employed or to increases in the productivity of the labor being employed. Increases in U.S. real GDP are mostly the result of increases in labor productivity. The increases in labor productivity can be attributed to technological progress, increases in the quantity of capital per worker, improvements in the education and training of workers, the exploitation of economies of scale, and improvements in the allocation of labor across different industries.

LO26.5 Explain why the trend rate of U.S. productivity growth has increased since the earlier 1973–1995 period.

Over long time periods, the growth of labor productivity underlies an economy's growth of real wages and its standard of living. U.S. productivity rose by 2.4 percent annually between 1995 and 2012, compared to 1.5 percent annually between 1973 and 1995.

This post-1995 increase in the average rate of productivity growth is based on (*a*) rapid technological change in the form of the microchip and information technology, (*b*) increasing returns and lower per-unit costs, and (*c*) heightened global competition that holds down prices.

The main sources of increasing returns in recent years are (*a*) the use of more specialized inputs as firms grow, (*b*) the spreading of development costs, (*c*) simultaneous consumption by consumers, (*d*) network effects, and (*e*) learning by doing. Increasing returns mean higher productivity and lower per-unit production costs.

LO26.6 Discuss differing perspectives as to whether growth is desirable and sustainable.

Skeptics wonder if the recent rise in the average rate of productivity growth is permanent, and suggest a wait-and-see approach. They point out that surges in productivity and real GDP growth have previously occurred but do not necessarily represent long-lived trends.

Critics of rapid growth say that it adds to environmental degradation, increases human stress, and exhausts the earth's finite supply of natural resources. Defenders of rapid growth say that it is the primary path to the rising living standards nearly universally desired by people, that it need not debase the environment, and that there are no indications that we are running out of resources. Growth is based on the expansion and application of human knowledge, which is limited only by human imagination.

TERMS AND CONCEPTS

economic growth	demand factor	economies of scale
real GDP per capita	efficiency factor	information technology
rule of 70	labor productivity	start-up firms
modern economic growth	labor-force participation rate	increasing returns
leader countries	growth accounting	network effects
follower countries	infrastructure	learning by doing
supply factors	human capital	

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. How is economic growth measured? Why is economic growth important? Why could the difference between a 2.5 percent and a 3 percent annual growth rate be of great significance over several decades? **LO26.1**

2. When and where did modern economic growth first happen? What are the major institutional factors that form the foundation for modern economic growth? What do they have in common? **LO26.2**

3. Why are some countries today much poorer than other countries? Are today's poor countries destined to always be poorer than today's rich countries? If so, explain why. If not, explain how today's poor countries can catch or even pass today's rich countries. **LO26.2**
4. What are the four supply factors of economic growth? What is the demand factor? What is the efficiency factor? Illustrate these factors in terms of the production possibilities curve. **LO26.3**
5. Suppose that Alpha and Omega have identically sized working-age populations but that total annual hours of work are much greater in Alpha than in Omega. Provide two possible reasons for this difference. **LO26.3**
6. What is growth accounting? To what extent have increases in U.S. real GDP resulted from more labor inputs? From greater labor productivity? Rearrange the following contributors to the growth of productivity in order of their quantitative importance: economies of scale, quantity of capital, improved resource allocation, education and training, and technological advance. **LO26.4**
7. True or False: If false, explain why. **LO26.4**
   a. Technological advance, which to date has played a relatively small role in U.S. economic growth, is destined to play a more important role in the future.
   b. Many public capital goods are complementary to private capital goods.
   c. Immigration has slowed economic growth in the United States.
8. Explain why there is such a close relationship between changes in a nation's rate of productivity growth and changes in its average real hourly wage. **LO26.5**
9. Relate each of the following to the recent increase in the trend rate of productivity growth: **LO26.5**
   a. Information technology.
   b. Increasing returns.
   c. Network effects.
   d. Global competition.
10. What, if any, are the benefits and costs of economic growth, particularly as measured by real GDP per capita? **LO26.6**
11. **LAST WORD** Would you expect a country with a total fertility rate of 2.7 to have a growing or a shrinking population over the long run? What about a country with a total fertility rate of 1.2? In 20 years, will America have more or fewer workers per retiree than it does today? Why does a falling inverse dependency ratio make it harder for real GDP to continue growing?

REVIEW QUESTIONS

1. If real GDP grows at 7 percent per year, then real GDP will double in approximately ________ years. **LO26.1**
   a. 70.
   b. 14.
   c. 10.
   d. 7.
2. In 1820 living standards in various places around the globe were ________ they are today. **LO26.2**
   a. More widely varying than.
   b. Just as widely varying as.
   c. Less widely varying than.
3. True or False: Countries that currently have low real GDPs per capita are destined to always have lower living standards than countries that currently have high real GDPs per capita. **LO26.2**
4. Identify each of the following situations as something that either promotes growth or retards growth. **LO26.2**
   a. Increasing corruption allows government officials to steal people's homes.
   b. A nation introduces patent laws for the first time.
   c. A court order shuts down all banks permanently.
   d. A poor country extends free public schooling from 8 years to 12 years.
   e. A nation adopts a free-trade policy.
   f. A formerly communist country adopts free markets.
5. Real GDP equals ________ times ________. **LO26.4**
   a. Average hours of work; quantity of capital.
   b. Average hours of work; allocative efficiency.
   c. Labor input; labor productivity.
   d. Natural resources; improvements in technology.
6. Suppose that just by doubling the amount of output that it produces each year, a firm's per-unit production costs fall by 30 percent. This is an example of: **LO26.4**
   a. Economies of scale.
   b. Improved resource allocation.
   c. Technological advance.
   d. The demand factor.
7. True or False: Computers and increased global competition have retarded economic growth in recent decades. **LO26.5**
8. Identify following arguments about economic growth as being either anti-growth or pro-growth. **LO26.6**
   a. Growth means worker burnout and frantic schedules.
   b. Rising incomes allow people to buy more education, medical care, and recreation.
   c. The Earth has only finite amounts of natural resources.
   d. We still have poverty, homelessness, and discrimination even in the richest countries.
   e. Richer countries spend more money protecting the environment.
   f. Natural resource prices have fallen rather than increased over time.

PROBLEMS

1. Suppose an economy's real GDP is \$30,000 in year 1 and \$31,200 in year 2. What is the growth rate of its real GDP? Assume that population is 100 in year 1 and 102 in year 2. What is the growth rate of real GDP per capita? **LO26.1**
2. What annual growth rate is needed for a country to double its output in 7 years? In 35 years? In 70 years? In 140 years? **LO26.1**
3. Assume that a "leader country" has real GDP per capita of \$40,000, whereas a "follower country" has real GDP per capita of \$20,000. Next suppose that the growth of real GDP per capita falls to zero percent in the leader country and rises to 7 percent in the follower country. If these rates continue for long periods of time, how many years will it take for the follower country to catch up to the living standard of the leader country? **LO26.2**
4. Refer to Figure 26.2 and assume that the values for points *a*, *b*, and *c* are \$10 billion, \$20 billion, and \$18 billion respectively. If the economy moves from point *a* to point *b* over a 10-year period, what must have been its annual rate of economic growth? If, instead, the economy was at point *c* at the end of the 10-year period, by what percentage did it fall short of its production capacity? **LO26.3**
5. Suppose that work hours in New Zombie are 200 in year 1 and productivity is \$8 per hour worked. What is New Zombie's real GDP? If work hours increase to 210 in year 2 and productivity rises to \$10 per hour, what is New Zombie's rate of economic growth? **LO26.4**
6. The per-unit cost of an item is its average total cost (= total cost/quantity). Suppose that a new cell phone application costs \$100,000 to develop and only \$0.50 per unit to deliver to each cell phone customer. What will be the per-unit cost of the application if it sells 100 units? 1,000 units? 1 million units? **LO26.5**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **27**

Business Cycles, Unemployment, and Inflation

Learning Objectives

LO27.1 **Describe the business cycle and its primary phases.**

LO27.2 **Illustrate how unemployment is measured and explain the different types of unemployment.**

LO27.3 **Explain how inflation is measured and distinguish between cost-push inflation and demand-pull inflation.**

LO27.4 **Relate how unanticipated inflation can redistribute real income.**

LO27.5 **Discuss how inflation may affect the economy's level of real output.**

As indicated in Chapter 26, the United States has experienced remarkable economic growth over time. But this growth has not been smooth, steady, and predictable from year to year. At various times the United States has experienced recessions, high unemployment rates, or high inflation rates. For example, U.S. unemployment rose by 8 million workers and the unemployment rate increased from 4.7 percent to 10.1 percent during the 2007–2009 recession. Other nations have also suffered high unemployment rates at times. As just one example, Spain's unemployment rate exceeded 26 percent in 2012. Also, inflation has occasionally plagued the United States and other nations. For instance, the U.S. inflation rate in 1980 was 13.5 percent. Zimbabwe's inflation soared to 26,000 percent in 2007!

Our goal in this chapter is to examine the concepts, terminology, and facts relating to macroeconomic instability. Specifically, we want to discuss the business cycle, unemployment, and inflation. The concepts discussed are extremely important for understanding subsequent chapters on economic theory and economic policy.

The Business Cycle

LO27.1 Describe the business cycle and its primary phases.

The long-run trend of the U.S. economy is one of economic growth, as stylized by the upsloping line labeled "Growth Trend" in Figure 27.1. But growth has been interrupted by periods of economic instability usually associated with **business cycles.** Business cycles are alternating rises and declines in the level of economic activity, sometimes over several years. Individual cycles (one "up" followed by one "down") vary substantially in duration and intensity.

ORIGIN OF THE IDEA

O27.1

Business cycles

Phases of the Business Cycle

Figure 27.1 shows the four phases of a generalized business cycle:

- At a **peak,** such as the middle peak shown in Figure 27.1, business activity has reached a temporary maximum. Here the economy is near or at full employment and the level of real output is at or very close to the economy's capacity. The price level is likely to rise during this phase.
- A **recession** is a period of decline in total output, income, and employment. This downturn, which lasts 6 months or more, is marked by the widespread contraction of business activity in many sectors of the economy. Along with declines in real GDP, significant increases in unemployment occur. Table 27.1

FIGURE 27.1 **The business cycle.** Economists distinguish four phases of the business cycle; the duration and strength of each phase may vary.

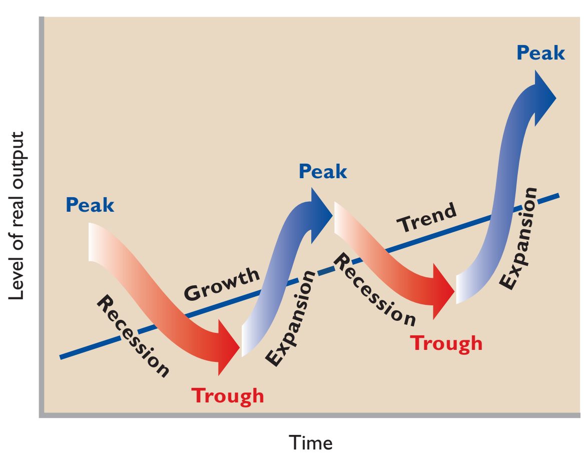


TABLE 27.1 **U.S. Recessions since 1950**

Period	Duration, Months	Depth (Decline in Real Output)
1953–54	10	−2.6%
1957–58	8	−3.7
1960–61	10	−1.1
1969–70	11	−0.2
1973–75	16	−3.2
1980	6	−2.2
1981–82	16	−2.9
1990–91	8	−1.4
2001	8	−0.4
2007–09	18	−4.3

Source: National Bureau of Economic Research, **www.nber.org**, Bureau of Economic Analysis, **www.bea.gov**, and Minneapolis Federal Reserve Bank, "The Recession and Recovery in Perspective," **www.minneapolisfed.gov**. Output data are in 2000 dollars.

documents the 10 recessions in the United States since 1950.

- In the **trough** of the recession or depression, output and employment "bottom out" at their lowest levels. The trough phase may be either short-lived or quite long.
- A recession is usually followed by a recovery and **expansion,** a period in which real GDP, income, and employment rise. At some point, the economy again approaches full employment. If spending then expands more rapidly than does production capacity, prices of nearly all goods and services will rise. In other words, inflation will occur.

Although business cycles all pass through the same phases, they vary greatly in duration and intensity. Many economists prefer to talk of business "fluctuations" rather than cycles because cycles imply regularity while fluctuations do not. The Great Depression of the 1930s resulted in a 27 percent decline in real GDP over a 3-year period in the United States and seriously impaired business activity for a decade. By comparison, the U.S. recessions detailed in Table 27.1 were less severe in both intensity and duration.

The Business Cycle Dating Committee of the National Bureau of Economic Research (NBER), a nonprofit economic research organization, declares the start and end of recessions in the United States. Citing evidence of declining real output and falling employment, the NBER officially declared that the latest recession began in December 2007. The NBER subsequently declared that the Great Recession ended in June 2009, 18 months after it began. In making this announcement, the NBER pointed out that its declaration was not a forecast for the future path of the economy.

Recessions, of course, occur in other countries, too. For example, nearly all industrial nations and many developing nations have suffered recessions in the past several years.

Causation: A First Glance

The long-run trend of the U.S. economy is expansion and growth. That is why the business cycles in Figure 27.1 are drawn against a trend of economic growth. A key issue in macroeconomics is why the economy sees business cycle fluctuations rather than slow, smooth growth. In terms of Figure 27.1, why does output move up and down rather than just staying on the smooth growth trend line?

Economists have developed several possible explanations. But before turning to them, recall that in Chapter 24 we explained that these theories are founded on the idea that fluctuations are driven by shocks—unexpected events that individuals and firms may have trouble adjusting to. Also recall that short-run price stickiness is widely believed to be a major factor preventing the economy from rapidly adjusting to shocks. With prices sticky in the short run, price changes cannot quickly equalize the quantities demanded of goods and services with their respective quantities supplied after a shock has happened. Instead, the economy is forced to respond to shocks in the short run primarily through changes in output and employment rather than through changes in prices.

Economists cite several possible general sources of shocks that can cause business cycles.

- ***Irregular innovation*** Significant new products or production methods, such as those associated with the railroad, automobile, computer, and Internet, can rapidly spread through the economy, sparking sizable increases in investment, consumption, output, and employment. After the economy has largely absorbed the new innovation, the economy may for a time slow down or possibly decline. Because such innovations occur irregularly and unexpectedly, they may contribute to the variability of economic activity.
- ***Productivity changes*** When productivity—output per unit of input—unexpectedly increases, the economy booms; when productivity unexpectedly decreases, the economy recedes. Such changes in productivity can result from unexpected changes in resource availability (of, say, oil or agricultural commodities) or from unexpected changes in the general rate of technological advance.
- ***Monetary factors*** Some economists see business cycles as purely monetary phenomena. When a nation's central bank shocks the economy by creating more money than people were expecting, an inflationary boom in output occurs. By contrast, printing less money than people were expecting triggers an output decline and, eventually, a price-level fall.
- ***Political events*** Unexpected political events, such as peace treaties, new wars, or the 9/11 terrorist attacks, can create economic opportunities or strains. In adjusting to these shocks, the economy may experience upswings or downswings.
- ***Financial instability*** Unexpected financial bubbles (rapid asset price increases) or bursts (abrupt asset price decreases) can spill over to the general economy by expanding or contracting lending, and boosting or eroding the confidence of consumers and businesses. Booms and busts in the rest of the economy may follow.

The severe recession of 2007–2009 was precipitated by a combination of excessive money and a financial frenzy that led to overvalued real estate and unsustainable mortgage debt. Institutions bundled this debt into new securities ("derivatives") that were sold to financial investors. Some of the investors, in turn, bought insurance against losses that might arise from the securities. As real estate prices plummeted and mortgage defaults unexpectedly rocketed, the securitization and insurance structure buckled and nearly collapsed. Credit markets froze, pessimism prevailed, and spending by businesses and households declined.

Whatever the source of economic shocks, most economists agree that the *immediate* cause of the large majority of cyclical changes in the levels of real output and employment is unexpected changes in the level of total spending. If total spending unexpectedly sinks and firms cannot lower prices, firms will find themselves selling fewer units of output (since with prices fixed, a decreased amount of spending implies fewer items purchased). Slower sales will cause firms to cut back on production. As they do, GDP will fall. And because fewer workers will be needed to produce less output, employment also will fall. The economy will contract and enter a recession.

By contrast, if the level of spending unexpectedly rises, output, employment, and incomes will rise. This is true because, with prices sticky, the increased spending will mean that consumers will be buying a larger volume of goods and services (since, with prices fixed, more spending means more items purchased). Firms will respond by increasing output. This will increase GDP. And because firms will need to hire more workers to produce the larger volume of output, employment also will increase. The economy will boom and enjoy an expansion. Eventually, as time passes and prices become more flexible, prices are also likely to rise as a result of the increased spending.

Cyclical Impact: Durables and Nondurables

Although the business cycle is felt everywhere in the economy, it affects different segments in different ways and to different degrees.

Firms and industries producing *capital goods* (for example, housing, commercial buildings, heavy equipment, and farm implements) and *consumer durables* (for example, automobiles, personal computers, and refrigerators) are affected most by the business cycle. Within limits, firms can postpone the purchase of capital goods. For instance, when the economy goes into recession, producers frequently delay the purchase of new equipment and the construction of new plants. The business outlook simply does not warrant increases in the stock of capital goods. In good times, capital goods are usually replaced before they depreciate completely. But when recession strikes, firms patch up their old equipment and make do. As a result, investment in capital goods declines sharply. Firms that have excess plant capacity may not even bother to replace all the capital that is depreciating. For them, net investment may be negative. The pattern is much the same for consumer durables such as automobiles and major appliances. When recession occurs and households must trim their budgets, purchases of these goods are often deferred. Families repair their old cars and appliances rather than buy new ones, and the firms producing these products suffer. (Of course, producers of capital goods and consumer durables also benefit most from expansions.)

In contrast, *service* industries and industries that produce *nondurable consumer goods* are somewhat insulated from the most severe effects of recession. People find it difficult to cut back on needed medical and legal services, for example. And a recession actually helps some service firms, such as pawnbrokers and law firms that specialize in bankruptcies. Nor are the purchases of many nondurable goods such as food and clothing easy to postpone. The quantity and quality of purchases of nondurables will decline, but not so much as will purchases of capital goods and consumer durables.

QUICK REVIEW 27.1

- The typical business cycle goes through four phases: peak, recession, trough, and expansion.
- Fluctuations in output and employment are caused by economic shocks combining with sticky prices.
- Sources of shocks that cause recessions include irregular innovation, productivity changes, monetary factors, political events, and financial instability.
- During a recession, industries that produce capital goods and consumer durables normally suffer greater output and employment declines than do service and nondurable consumer goods industries.

Unemployment

LO27.2 Illustrate how unemployment is measured and explain the different types of unemployment.

Two problems that arise over the course of the business cycle are unemployment and inflation. Let's look at unemployment first.

Measurement of Unemployment

The U.S. Bureau of Labor Statistics (BLS) conducts a nationwide random survey of some 60,000 households each month to determine who is employed and who is not employed. In a series of questions, it asks which members of the household are working, unemployed and looking for work, not looking for work, and so on. From the answers, it determines an unemployment rate for the entire nation.

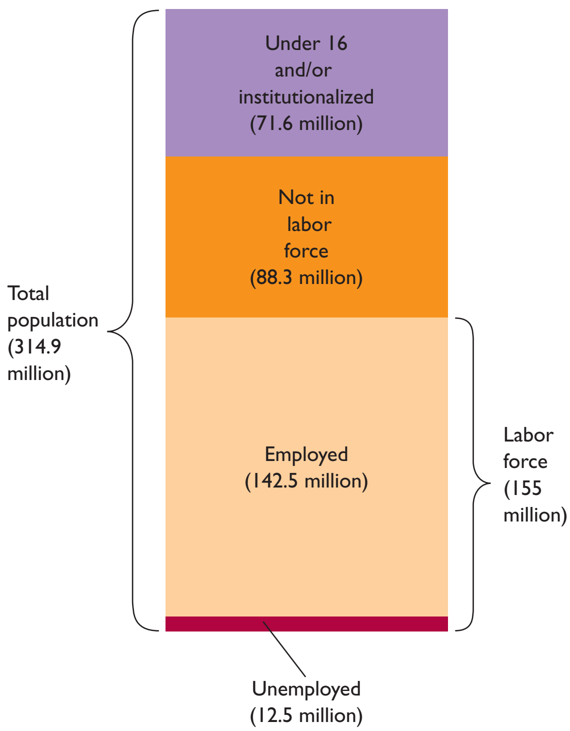
Figure 27.2 helps explain the mathematics. The BLS divides the total U.S. population into three groups. One group is made up of people under 16 years of age and people who are institutionalized, for example, in mental hospitals or correctional institutions. Such people are not considered potential members of the labor force.

A second group, labeled "Not in labor force," is composed of adults who are potential workers but are not employed and are not seeking work. For example, they are stay-at-home parents, full-time students, or retirees.

The third group is the **labor force,** which constituted slightly more than 50 percent of the total population in 2009. The labor force consists of people who are able and willing to work. Both those who are employed and those who are unemployed but actively seeking work are counted as being in the labor force. The **unemployment rate** is the percentage of the labor force unemployed:

$$\text{Unemployment rate} = \frac{\text{unemployed}}{\text{labor force}} \times 100$$

FIGURE 27.2 The U.S. labor force, employment, and unemployment, 2012.* The labor force consists of persons 16 years of age or older who are not in institutions and who are (1) employed or (2) unemployed but seeking employment.



*Civilian labor-force data, which excludes military employment.

Source: Bureau of Labor Statistics, **www.bls.gov**.

The statistics underlying the rounded numbers in Figure 27.2 show that in 2012 the unemployment rate averaged

$$\frac{12{,}506{,}000}{154{,}975{,}000} \times 100 = 8.1\%$$

WORKED PROBLEMS

W27.1
Unemployment rate

Unemployment rates for selected years appear on the inside covers of this book.

Despite the use of scientific sampling and interviewing techniques, the data collected in this survey are subject to criticism:

- ***Part-time employment*** The BLS lists all part-time workers as fully employed. In 2012 about 26 million people worked part-time as a result of personal choice. But another 8 million part-time workers either wanted to work full-time and could not find suitable full-time work or worked fewer hours because of a temporary slack in consumer demand. These last two groups were, in effect, partially employed and partially unemployed. By counting them as fully employed, say critics, the official BLS data understate the unemployment rate.
- ***Discouraged workers*** You must be actively seeking work in order to be counted as unemployed. An unemployed individual who is not actively seeking employment is classified as "not in the labor force." The problem is that many workers, after unsuccessfully seeking employment for a time, become discouraged and drop out of the labor force. The number of such **discouraged workers** was roughly 909,000 in 2012, up from 396,000 in 2007. By not counting discouraged workers as unemployed, say critics, the official BLS data understate the unemployment problem.

Types of Unemployment

There are three *types* of unemployment: frictional, structural, and cyclical.

Frictional Unemployment At any given time some workers are "between jobs." Some of them will be moving voluntarily from one job to another. Others will have been fired and will be seeking reemployment. Still others will have been laid off temporarily because of seasonal demand. In addition to those between jobs, many young workers will be searching for their first jobs.

As these unemployed people find jobs or are called back from temporary layoffs, other job seekers and laid-off workers will replace them in the "unemployment pool." It is important to keep in mind that while the pool itself persists because there are always newly unemployed workers flowing into it, most workers do *not* stay in the unemployment pool for very long. Indeed, when the economy is strong, the majority of unemployed workers find new jobs within a couple of months. One should be careful not to make the mistake of confusing the permanence of the pool itself with the false idea that the pool's membership is permanent, too. On the other hand, there are workers who do remain unemployed and in the pool for very long periods of time—sometimes for many years. As we discuss the different types of unemployment below, notice that certain types tend to be transitory while others are associated with much longer spells of unemployment.

Economists use the term **frictional unemployment**—consisting of *search unemployment* and *wait unemployment*—for workers who are either searching for jobs or waiting to take jobs in the near future. The word "frictional" implies

CONSIDER THIS . . .

Downwardly Sticky Wages and Unemployment

Labor markets have an important quirk that helps to explain why unemployment goes up so much during a recession.

The quirk is that wages are flexible upward but sticky downward.

On the one hand, workers are perfectly happy to accept wage increases. So when the economy is booming and firms start bidding for the limited supply of labor, wages rise—often quite rapidly.

On the other hand, workers deeply resent pay cuts. So if the economy goes into a recession and firms need to reduce labor costs, managers almost never cut wages because doing so would only lead to disgruntled employees, low productivity, and—in extreme cases—workers stealing supplies or actively sabotaging their own firms.

Instead, managers usually opt for layoffs. The workers who are let go obviously don't like being unemployed. But those who remain get to keep their old wages and, consequently, keep on being as productive and cooperative as they were before.

This preference that firms show for layoffs over wage cuts results in downwardly sticky wages and an informal price floor that helps to explain why unemployment goes up so much during a recession. The problem is that when the demand for labor falls during a recession, the informal price floor prevents wages from falling. As a result, there is no way for falling wages to help entice at least some firms to hire a few more workers. Thus, when a recession hits, employment falls more precipitously than it would if wages were downwardly flexible and falling wages could help to increase hiring.

that the labor market does not operate perfectly and instantaneously (without friction) in matching workers and jobs.

Frictional unemployment is inevitable and, at least in part, desirable. Many workers who are voluntarily between jobs are moving from low-paying, low-productivity jobs to higher-paying, higher-productivity positions. That means greater income for the workers, a better allocation of labor resources, and a larger real GDP for the economy.

Structural Unemployment Frictional unemployment blurs into a category called **structural unemployment.** Here, economists use "structural" in the sense of "compositional." Changes over time in consumer demand and in technology alter the "structure" of the total demand for labor, both occupationally and geographically.

Occupationally, the demand for certain skills (for example, sewing clothes or working on farms) may decline or even vanish. The demand for other skills (for example, designing software or maintaining computer systems) will intensify. Unemployment results because the composition of the labor force does not respond immediately or completely to the new structure of job opportunities. Workers who find that their skills and experience have become obsolete or unneeded thus find that they have no marketable talents. They are structurally unemployed until they adapt or develop skills that employers want.

Geographically, the demand for labor also changes over time. An example: the migration of industry and thus of employment opportunities from the Snowbelt to the Sunbelt over the past few decades. Another example is the movement of jobs from inner-city factories to suburban industrial parks. And a final example is the so-called *offshoring* of jobs that occurs when the demand for a particular type of labor shifts from domestic firms to foreign firms. As job opportunities shift from one place to another, some workers become structurally unemployed.

The distinction between frictional and structural unemployment is hazy at best. The key difference is that *frictionally* unemployed workers have marketable skills and either live in areas where jobs exist or are able to move to areas where they do. *Structurally* unemployed workers find it hard to obtain new jobs without retraining, gaining additional education, or relocating. Frictional unemployment is short-term; structural unemployment is more likely to be long-term and consequently more serious.

Cyclical Unemployment Unemployment that is caused by a decline in total spending is called **cyclical unemployment** and typically begins in the recession phase of the business cycle. As the demand for goods and services decreases, employment falls and unemployment rises. Cyclical unemployment results from insufficient demand for goods and services. The 25 percent unemployment rate in the depth of the Great Depression in 1933 reflected mainly cyclical unemployment, as did significant parts of the 9.7 percent unemployment rate in 1982, the 7.5 percent rate in 1992, the 5.8 percent rate in 2002, and the 9.3 percent rate in 2009.

Cyclical unemployment is a very serious problem when it occurs. We will say more about its high costs later, but first we need to define "full employment."

Definition of Full Employment

Because frictional and structural unemployment are largely unavoidable in a dynamic economy, *full employment* is something less than 100 percent employment of the labor force. Economists say that the economy is "fully employed" when it is experiencing only frictional and structural unemployment. That is, full employment occurs when there is no cyclical unemployment.

Economists describe the unemployment rate that is consistent with full employment as the **full-employment rate of unemployment,** or the **natural rate of unemployment (NRU).** At the NRU, the economy is said to be producing its **potential output.** This is the real GDP that occurs when the economy is "fully employed."

Note that a fully employed economy does not mean zero unemployment. Even when the economy is fully employed, the NRU is some positive percentage because it takes time for frictionally unemployed job seekers to find open jobs they can fill. Also, it takes time for the structurally unemployed to achieve the skills and geographic relocation needed for reemployment.

"Natural" does not mean, however, that the economy will always operate at this rate and thus realize its potential output. When cyclical unemployment occurs, the economy has much more unemployment than that which would occur at the NRU. Moreover, the economy can operate for a while at an unemployment rate *below* the NRU. At times, the demand for labor may be so great that firms take a stronger initiative to hire and train the structurally unemployed. Also, some parents, teenagers, college students, and retirees who were casually looking for just the right part-time or full-time jobs may quickly find them. Thus the unemployment rate temporarily falls below the natural rate.

Also, the NRU can vary over time as demographic factors, job-search methods, and public policies change. In the 1980s, the NRU was about 6 percent. Today, it is 5 to 6 percent.

Economic Cost of Unemployment

Unemployment that is excessive involves great economic and social costs.

GDP Gap and Okun's Law The basic economic cost of unemployment is forgone output. When the economy fails to create enough jobs for all who are able and willing to work, potential production of goods and services is irretrievably lost. In terms of Chapter 1's analysis, unemployment above the natural rate means that society is operating at some point inside its production possibilities curve. Economists call this sacrifice of output a **GDP gap**—the difference between actual and potential GDP. That is:

$$\text{GDP gap} = \text{actual GDP} - \text{potential GDP}$$

The GDP gap can be either negative (actual GDP $<$ potential GDP) or positive (actual GDP $>$ potential GDP). In the case of unemployment above the natural rate, it is negative because actual GDP falls short of potential GDP.

Potential GDP is determined by assuming that the natural rate of unemployment prevails. The growth of potential GDP is simply projected forward on the basis of the economy's "normal" growth rate of real GDP. Figure 27.3 shows the GDP gap for recent years in the United States. It also indicates the close correlation between the actual unemployment rate (Figure 27.3b) and the GDP gap (Figure 27.3a). The higher the unemployment rate, the larger is the GDP gap.

Macroeconomist Arthur Okun was the first to quantify the relationship between the unemployment rate and the GDP gap. **Okun's law** indicates that for every 1 percentage point by which the actual unemployment rate exceeds the natural rate, a negative GDP gap of about 2 percent occurs. With this information, we can calculate the absolute loss of output associated with any above-natural unemployment rate. For example, in 2009 the unemployment rate was 9.3 percent, or 4.3 percentage points above that period's 5.0 percent natural rate of unemployment. Multiplying this 4.3 percent by Okun's 2 indicates that 2009's GDP gap was 8.6 percent of potential GDP (in real terms). By applying this 8.6 percent loss to 2009's potential GDP of $13,894 billion, we find that the economy sacrificed $1,195 billion of real output because the natural rate of unemployment was not achieved.

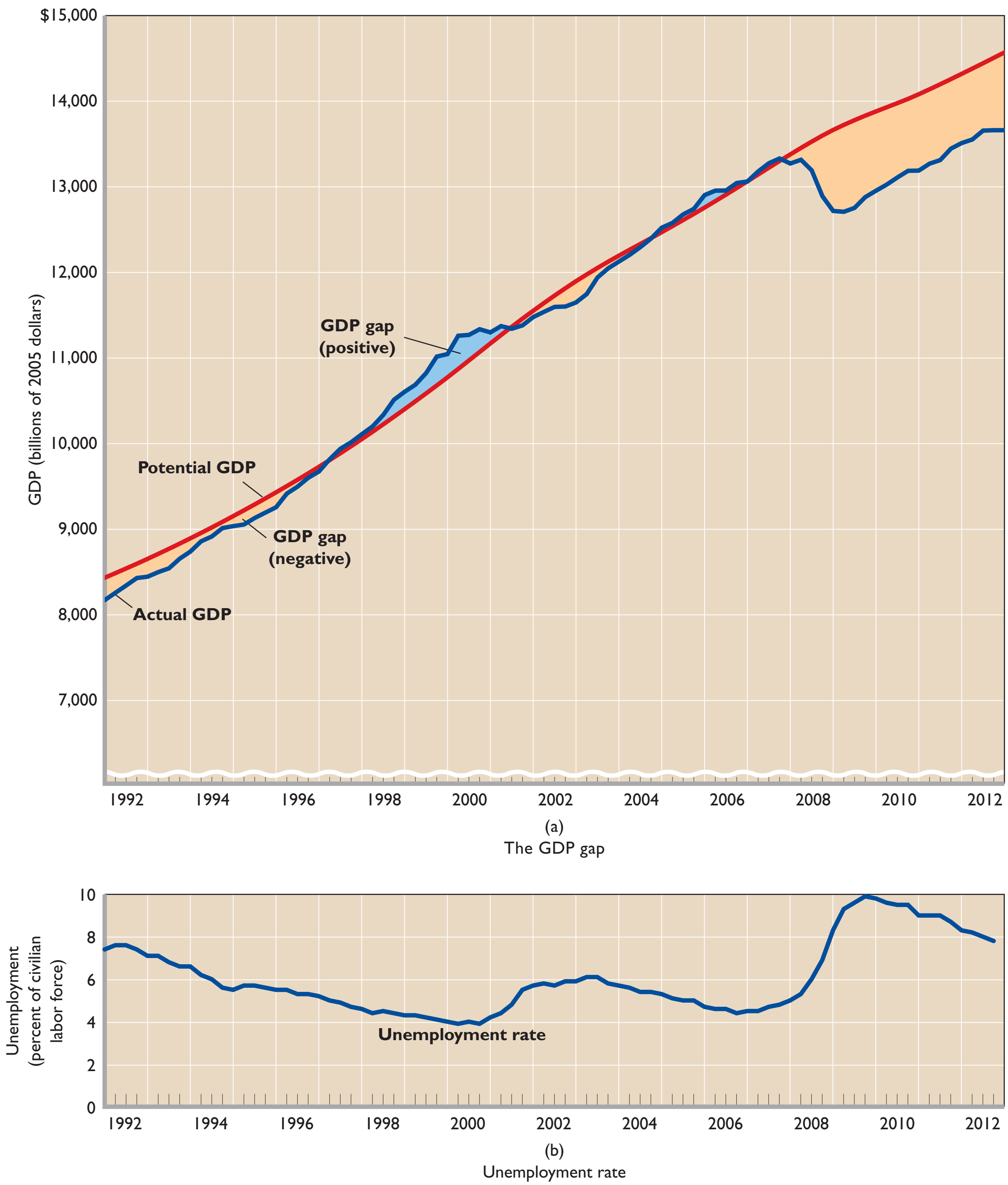
WORKED PROBLEMS

W27.2
Okun's law

As you can see in Figure 27.3, sometimes the economy's actual output will exceed its potential or full-employment output. Figure 27.3 reveals that an economic expansion in 1999 and 2000, for example, caused actual GDP to exceed potential GDP in those years. There was a positive GDP gap in 1999 and 2000. Actual GDP for a time can exceed potential GDP, but positive GDP gaps create inflationary pressures and cannot be sustained indefinitely.

Unequal Burdens An increase in the unemployment rate from 5 to, say, 9 or 10 percent might be more tolerable to society if every worker's hours of work and wage

FIGURE 27.3 Actual and potential real GDP and the unemployment rate. (a) The difference between actual and potential GDP is the GDP gap. A negative GDP gap measures the output the economy sacrifices when actual GDP falls short of potential GDP. A positive GDP gap indicates that actual GDP is above potential GDP. (b) A high unemployment rate means a large GDP gap (negative), and a low unemployment rate means a small or even positive GDP gap.



Source: Congressional Budget Office, **www.cbo.gov**; Bureau of Economic Analysis, **www.bea.gov**; and the Bureau of Labor Statistics, **www.bls.gov**. Note that the data for actual real GDP and potential real GDP above differ from the real GDP data in the previous two chapters. The data above are in 2005 dollars, not 2009 dollars, and do not reflect the 2013 redefinition of investment expenditures in the National Income and Product Accounts to include research and development spending.

TABLE 27.2 Unemployment Rates by Demographic Group: Full Employment Year (2007) and Recession Year (2009)*

	Unemployment Rate	
Demographic Group	**2007**	**2009**
Overall	4.6%	9.3%
Occupation:		
Managerial and professional	2.1	4.6
Construction and extraction	7.6	19.7
Age:		
16–19	15.7	24.3
African American, 16–19	29.4	39.5
White, 16–19	13.9	21.8
Male, 20+	4.1	9.6
Female, 20+	4.0	7.5
Race and ethnicity:		
African American	8.3	14.8
Hispanic	5.6	12.1
White	4.1	8.5
Gender:		
Women	4.5	8.1
Men	4.7	10.3
Education:†		
Less than high school diploma	7.1	14.6
High school diploma only	4.4	9.7
College degree or more	2.0	4.6
Duration:		
15 or more weeks	1.5	4.7

*Civilian labor-force data.

†People age 25 or over.

Source: Economic Report of the President; Bureau of Labor Statistics, **www.bls.gov**; Census Bureau, **www.census.gov**.

income were reduced proportionally. But this is not the case. Part of the burden of unemployment is that its cost is unequally distributed.

Table 27.2 examines unemployment rates for various labor market groups for 2 periods. In 2007, the economy achieved full employment, with a 4.6 percent unemployment rate. The economy receded in December 2007 and two years later was feeling the full unemployment impact of the Great Recession. By observing the large variance in unemployment rates for the different groups within each period and comparing the rates between the 2 periods, we can generalize as follows:

- ***Occupation*** Workers in lower-skilled occupations (for example, laborers) have higher unemployment rates than workers in higher-skilled occupations (for example, professionals). Lower-skilled workers have more and longer spells of structural unemployment than higher-skilled workers. They also are less likely to be self-employed than are higher-skilled workers. Moreover, lower-skilled workers usually bear the brunt of recessions. Manufacturing, construction, and mining tend to be particularly hard-hit, and businesses generally retain most of their higher-skilled workers, in whom they have invested the expense of training.
- ***Age*** Teenagers have much higher unemployment rates than adults. Teenagers have lower skill levels, quit their jobs more frequently, are more frequently fired, and have less geographic mobility than adults. Many unemployed teenagers are new in the labor market, searching for their first jobs. Male African-American teenagers, in particular, have very high unemployment rates. The unemployment rate for all teenagers rises during recessions.
- ***Race and ethnicity*** The unemployment rates for African Americans and Hispanics are higher than that for whites. The causes of the higher rates include lower rates of educational attainment, greater concentration in lower-skilled occupations, and discrimination in the labor market. In general, the unemployment rate for African Americans is twice that of whites and rises by more percentage points than for whites during recessions.
- ***Gender*** The unemployment rates for men and women normally are very similar. But in the recent recession, the unemployment rate for men significantly exceeded that for women.
- ***Education*** Less-educated workers, on average, have higher unemployment rates than workers with more education. Less education is usually associated with lower-skilled, less-permanent jobs; more time between jobs; and jobs that are more vulnerable to cyclical layoff.
- ***Duration*** The number of persons unemployed for long periods—15 weeks or more—as a percentage of the labor force is much lower than the overall unemployment rate. But that percentage rises significantly during recessions. Notice from Table 27.2 that it rose from 1.5 percent of the labor force in 2007 to 4.7 percent in 2009.

Noneconomic Costs

Severe cyclical unemployment is more than an economic malady; it is a social catastrophe. Unemployment means idleness. And idleness means loss of skills, loss of self-respect, plummeting morale, family disintegration, and sociopolitical unrest. Widespread joblessness increases poverty, heightens racial and ethnic tensions, and reduces hope for material advancement.

History demonstrates that severe unemployment can lead to rapid and sometimes violent social and political change. Witness Hitler's ascent to power against a background of unemployment in Germany. Furthermore, relatively high unemployment among some racial and ethnic minorities has contributed to the unrest and violence that has periodically plagued some cities in the United States and abroad. At the individual level, research links increases in suicide, homicide, fatal heart attacks and strokes, and mental illness to high unemployment.

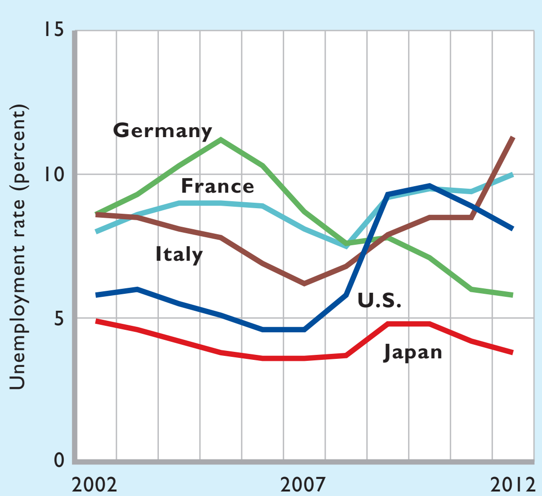
International Comparisons

Unemployment rates differ greatly among nations at any given time. One reason is that nations have different natural rates of unemployment. Another is that nations may be in different phases of their business cycles. Global Perspective 27.1 shows unemployment rates for five industrialized nations for the years 2002 through 2012. Between 2002 and 2008, the U.S. unemployment rate was considerably lower than the rates in Italy, France, and Germany. But during the Great Recession, U.S. unemployment spiked to the highest level among the five countries.

GLOBAL PERSPECTIVE 27.1

Unemployment Rates in Five Industrial Nations, 2002–2012

Compared with Italy, France, and Germany, the United States had a relatively low unemployment rate until the start of the 2007–2009 Great Recession, when the U.S. rate shot up to become the highest among the five nations.



Source: Bureau of Labor Statistics, **www.bls.gov**. Based on U.S. unemployment concepts.

QUICK REVIEW 27.2

- Unemployment is of three general types: frictional, structural, and cyclical.
- The natural unemployment rate (frictional plus structural) is presently 5 to 6 percent in the United States.
- A positive GDP gap occurs when actual GDP exceeds potential GDP; a negative GDP gap occurs when actual GDP falls short of potential GDP.
- Society loses real GDP when cyclical unemployment occurs; according to Okun's law, for each 1 percentage point of unemployment above the natural rate, the U.S. economy suffers a 2 percent decline in real GDP below its potential GDP.
- Lower-skilled workers, teenagers, African Americans and Hispanics, and less-educated workers bear a disproportionate burden of unemployment.

Inflation

LO27.3 Explain how inflation is measured and distinguish between cost-push inflation and demand-pull inflation.

We now turn to inflation, another aspect of macroeconomic instability. The problems inflation poses are subtler than those posed by unemployment.

Meaning of Inflation

Inflation is a rise in the general level of prices. When inflation occurs, each dollar of income will buy fewer goods and services than before. Inflation reduces the "purchasing power" of money. But inflation does not mean that *all* prices are rising. Even during periods of rapid inflation, some prices may be relatively constant and others may even fall. For example, although the United States experienced high rates of inflation in the 1970s and early 1980s, the prices of video recorders, digital watches, and personal computers declined.

Measurement of Inflation

The main measure of inflation in the United States is the **Consumer Price Index (CPI),** compiled by the BLS. The government uses this index to report inflation rates each month and each year. It also uses the CPI to adjust Social Security benefits and income tax brackets for inflation. The CPI reports the price of a "market basket" of some 300 consumer goods and services that are purchased by a typical urban consumer. (The GDP price index of Chapter 25 is a much broader measure of inflation since it includes not only consumer goods and services but also capital goods, goods and services purchased by government, and goods and services that enter world trade.)

The composition of the market basket for the CPI is based on spending patterns of urban consumers in a specific period, presently 2009–2010. The BLS updates the composition of the market basket every 2 years so that it reflects the most recent patterns of consumer purchases and captures the inflation that consumers are currently experiencing. The BLS arbitrarily sets the CPI equal to 100 for 1982–1984. So the CPI for any particular year is found as follows:

$$\text{CPI} = \frac{\text{price of the most recent market basket in the particular year}}{\text{price estimate of the market basket in 1982–1984}} \times 100$$

The rate of inflation is equal to the percentage growth of CPI from one year to the next. For example, the CPI was 207.3 in 2007, up from 201.6 in 2006. So the rate of inflation for 2007 is calculated as follows:

$$\text{Rate of inflation} = \frac{207.3 - 201.6}{201.6} \times 100 = 2.8\%$$

In rare cases, the CPI declines from one year to the next. For example, the CPI fell from 215.3 in 2008 to 214.5 in 2009. The rate of inflation for 2009 therefore was −0.4 percent. Such price level declines are called **deflation.**

In Chapter 26, we discussed the mathematical approximation called *the rule of 70*, which tells us that we can find the number of years it will take for some measure to double, given its annual percentage increase, by dividing that percentage increase into the number 70. So a 3 percent annual rate of inflation will double the price level in about 23 (= 70 ÷ 3) years. Inflation of 8 percent per year will double the price level in about 9 (= 70 ÷ 8) years.

Facts of Inflation

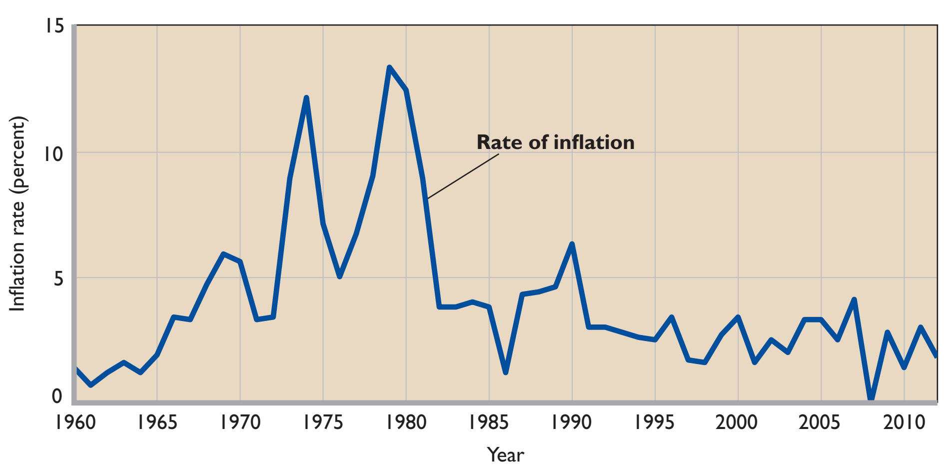
Figure 27.4 shows December-to-December rates of annual inflation in the United States between 1960 and 2011. Observe that inflation reached double-digit rates in the 1970s and early 1980s but has since declined and has been relatively mild recently.

In recent years U.S. inflation has been neither unusually high nor low relative to inflation in several other industrial countries (see Global Perspective 27.2). Some nations (not shown) have had double-digit or even higher annual rates of inflation in recent years. In 2009, for example, the annual inflation rate in the Democratic Republic of Congo was 46 percent; Eritrea, 35 percent; Afghanistan, 31 percent; and Venezuela, 27 percent. Zimbabwe's inflation rate was 14.9 billion percent in 2008 before Zimbabwe did away with its existing currency.

Types of Inflation

Nearly all prices in the economy are set by supply and demand. Consequently, if the economy is experiencing inflation and the overall level of prices is rising, we need to look for an explanation in terms of supply and demand.

FIGURE 27.4 Annual inflation rates in the United States, 1960–2011 (December-to-December changes in the CPI). The major periods of inflation in the United States in the past 51 years were in the 1970s and 1980s.

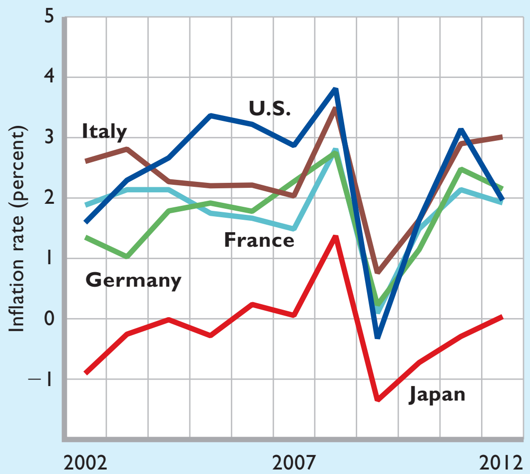


Source: Bureau of Labor Statistics, **www.bls.gov**.

GLOBAL PERSPECTIVE 27.2

Inflation Rates in Five Industrial Nations, 2002–2012

Inflation rates in the United States in recent years were neither extraordinarily high nor extraordinarily low relative to rates in other industrial nations.



Source: World Economic Outlook Database, April 2013, International Monetary Fund, **www.imf.org**.

Demand-Pull Inflation Usually, increases in the price level are caused by an excess of total spending beyond the economy's capacity to produce. Where inflation is rapid and sustained, the cause invariably is an overissuance of money by the central bank (the Federal Reserve in the United States). When resources are already fully employed, the business sector cannot respond to excess demand by expanding output. So the excess demand bids up the prices of the limited output, producing **demand-pull inflation.** The essence of this type of inflation is "too much spending chasing too few goods."

Cost-Push Inflation Inflation also may arise on the supply, or cost, side of the economy. During some periods in U.S. economic history, including the mid-1970s, the price level increased even though total spending was not excessive. These were periods when output and employment were both *declining* (evidence that total spending was not excessive) while the general price level was *rising*.

The theory of **cost-push inflation** explains rising prices in terms of factors that raise **per-unit production costs** at each level of spending. A per-unit production cost is the average cost of a particular level of output. This average cost is found by dividing the total cost of all resource inputs by the amount of output produced. That is,

$$\text{Per-unit production cost} = \frac{\text{total input cost}}{\text{units of output}}$$

Rising per-unit production costs squeeze profits and reduce the amount of output firms are willing to supply at the existing price level. As a result, the economy's supply of goods and services declines and the price level rises. In this scenario, costs are *pushing* the price level upward, whereas in demand-pull inflation demand is *pulling* it upward.

The major source of cost-push inflation has been so-called *supply shocks*. Specifically, abrupt increases in the costs of raw materials or energy inputs have on occasion driven up per-unit production costs and thus product prices. The rocketing prices of imported oil in 1973–1974 and again in 1979–1980 are good illustrations. As energy prices surged upward during these periods, the costs of producing and transporting virtually every product in the economy rose. Cost-push inflation ensued.

Complexities

The real world is more complex than the distinction between demand-pull and cost-push inflation suggests. It is difficult to distinguish between demand-pull inflation and cost-push inflation unless the original source of inflation is known. For example, suppose a significant increase in total spending occurs in a fully employed economy, causing demand-pull inflation. But as the demand-pull stimulus works its way through various product and resource markets, individual firms find their wage costs, material costs, and fuel prices rising. From their perspective they must raise their prices because production costs (someone else's prices) have risen. Although this inflation is clearly demand-pull in origin, it may mistakenly appear to be cost-push inflation to business firms and to government. Without proper identification of the source of the inflation, government and the Federal Reserve may be slow to undertake policies to reduce excessive total spending.

Another complexity is that cost-push inflation and demand-pull inflation differ in their sustainability. Demand-pull inflation will continue as long as there is excess total spending. Cost-push inflation is automatically self-limiting; it will die out by itself. Increased per-unit costs will reduce supply, and this means lower real output and employment. Those decreases will constrain further per-unit cost increases. In other words, cost-push inflation generates a recession. And in a recession, households and businesses concentrate on keeping their resources employed, not on pushing up the prices of those resources.

Core Inflation

Another complication relating to inflation (regardless of type) is noteworthy. Some price-flexible items within the consumer price index—particularly, food and energy—experience rapid changes in supply and demand and therefore considerable price volatility from month to month and year to year. For example, the prices of grain, fruit, vegetables, and livestock sometimes move rapidly in one direction or the other, leading to sizable changes in the prices of food items such as bread, oranges, lettuce, and beef. Also, energy items such as gasoline and natural gas can rise or fall rapidly from period to period. These ups and downs of food and energy prices usually are temporary and often cancel each other out over longer periods.

In tracking inflation, policymakers want to avoid being misled by rapid but temporary price changes that may distort the inflation picture. They mainly are interested in how rapidly the prices of the typically more stable components of the CPI are rising. By stripping volatile food and energy prices from the CPI, policymakers isolate so-called **core inflation,** the underlying increases in the CPI after volatile food and energy prices are removed.

If core inflation is low and stable, policymakers may be satisfied with current policy even though changes in the overall CPI index may be suggesting a rising rate of inflation. But policymakers become greatly concerned when core inflation is high and rising and take deliberate measures to try to halt it. We discuss these policies in later chapters.

CONSIDER THIS . . .

Clipping Coins

Some interesting early episodes of demand-pull inflation occurred in Europe from the ninth century to the fifteenth century under feudalism. In that economic system, *lords* (or *princes*) ruled individual fiefdoms and their *vassals* (or *peasants*) worked the fields. The peasants initially paid parts of their harvest as taxes to the princes. Later, when the princes began issuing "coins of the realm," peasants began paying their taxes with gold coins.

Some princes soon discovered a way to transfer purchasing power from their vassals to themselves without explicitly increasing taxes. As coins came into the treasury, princes clipped off parts of the gold coins, making them slightly smaller. From the clippings they minted new coins and used them to buy more goods for themselves.

This practice of clipping coins was a subtle form of taxation. The quantity of goods being produced in the fiefdom remained the same, but the number of gold coins increased. With "too much money chasing too few goods," inflation occurred. Each gold coin earned by the peasants therefore had less purchasing power than previously because prices were higher. The increase of the money supply shifted purchasing power away from the peasants and toward the princes just as surely as if the princes had increased taxation of the peasants.

In more recent eras some dictators have simply printed money to buy more goods for themselves, their relatives, and their key loyalists. These dictators, too, have levied hidden taxes on their population by creating inflation.

The moral of the story is quite simple: A society that values price-level stability should not entrust the control of its money supply to people who benefit from inflation.

QUICK REVIEW 27.3

- Inflation is a rising general level of prices and is measured as a percentage change in a price index such as the CPI; deflation is a decline in the general level of prices.
- For the past several years, the U.S. inflation rate has been within the general range of the rates of other advanced industrial nations and far below the rates experienced by some nations.
- Demand-pull inflation occurs when total spending exceeds the economy's ability to provide goods and services at the existing price level; total spending *pulls* the price level upward.
- Cost-push inflation occurs when factors such as rapid increases in the prices of imported raw materials drive up per-unit production costs at each level of output; higher costs *push* the price level upward.
- Core inflation is the underlying inflation rate after volatile food and energy prices have been removed.

Redistribution Effects of Inflation

LO27.4 Relate how unanticipated inflation can redistribute real income.

Inflation redistributes real income. This redistribution helps some people and hurts some others while leaving many people largely unaffected. Who gets hurt? Who benefits? Before we can answer, we need some terminology.

Nominal and Real Income There is a difference between money (or nominal) income and real income.

Nominal income is the number of dollars received as wages, rent, interest, or profit. **Real income** is a measure of the amount of goods and services nominal income can buy; it is the purchasing power of nominal income, or income adjusted for inflation. That is,

$$\text{Real income} = \frac{\text{nominal income}}{\text{price index (in hundredths)}}$$

Inflation need not alter an economy's overall real income—its total purchasing power. It is evident from the above equation that real income will remain the same when nominal income rises at the same percentage rate as does the price index.

But when inflation occurs, not everyone's nominal income rises at the same pace as the price level. Therein lies the potential for redistribution of real income from some to others. If the change in the price level differs from the change in a person's nominal income, his or her real income will be affected. The following approximation (shown by the ≅ sign) tells us roughly how much real income will change:

$$\begin{array}{l} \text{Percentage} \\ \text{change in} \\ \text{real income} \end{array} \cong \begin{array}{l} \text{percentage} \\ \text{change in} \\ \text{nominal income} \end{array} - \begin{array}{l} \text{percentage} \\ \text{change in} \\ \text{price level} \end{array}$$

For example, suppose that the price level rises by 6 percent in some period. If Bob's nominal income rises by 6 percent, his real income will *remain unchanged*. But if his nominal income instead rises by 10 percent, his real income will *increase* by about 4 percent. And if Bob's nominal income rises by only 2 percent, his real income will *decline* by about 4 percent.[1]

WORKED PROBLEMS

W27.3

Nominal and real income

Anticipations The redistribution effects of inflation depend upon whether or not it is expected. We will first discuss situations involving **unanticipated inflation.** As you will see, these cause real income and wealth to be redistributed, harming some and benefiting others. We will then discuss situations involving **anticipated inflation.** These are situations in which people see an inflation coming in advance. With the ability to plan ahead, people are able to avoid or lessen the redistribution effects associated with inflation.

[1]A more precise calculation uses our equation for real income. In our first illustration above, if nominal income rises by 10 percent from \$100 to \$110 and the price level (index) rises by 6 percent from 100 to 106, then real income has increased as follows:

$$\frac{\$110}{1.06} = \$103.77$$

The 4 percent increase in real income shown by the simple formula in the text is a reasonable approximation of the 3.77 percent yielded by our more precise formula.

Who Is Hurt by Inflation?

Unanticipated inflation hurts fixed-income recipients, savers, and creditors. It redistributes real income away from them and toward others.

Fixed-Income Receivers People whose incomes are fixed see their real incomes fall when inflation occurs. The classic case is the elderly couple living on a private pension or annuity that provides a fixed amount of nominal income each month. They may have retired in, say, 1993 on what

CONSIDER THIS . . .

Could a Little Inflation Help Reduce Unemployment?

Economists have debated whether a little inflation—say two or three percent per year—might help to reduce the unemployment rate during recessions.

Proponents argue that a little inflation might have this beneficial effect by boosting firms' profits and their demand for labor. Their argument goes like this. If wages and other costs were to remained fixed while inflation increased the prices at which firms could sell their output, firms would see their profitability increase. That in turn would cause firms to want to hire more workers.

The economists who disagree argue that it is implausible to assume that wages and other costs would remain fixed while inflation drives up the price of output. They point out that wages and other costs may well rise as fast or possibly even faster than output prices rise. If so, firms would not see any increase in their profitability—and thus they would not see any reason to hire more workers.

In addition, the economists who disagree also point out that even if inflation did lower unemployment, it would do so at the cost of lowering real wages. That's because if wages stay fixed while output prices rise, workers' fixed paychecks would only be able to purchase a smaller amount of goods and services. So while more workers might have jobs, those with jobs would have a lower standard of living.

appeared to be an adequate pension. However, by 2009 they would have discovered that inflation had cut the annual purchasing power of that pension—their real income—by one-third.

Similarly, landlords who receive lease payments of fixed dollar amounts will be hurt by inflation as they receive dollars of declining value over time. Likewise, public sector workers whose incomes are dictated by fixed pay schedules may suffer from inflation. The fixed "steps" (the upward yearly increases) in their pay schedules may not keep up with inflation. Minimum-wage workers and families living on fixed welfare incomes also will be hurt by inflation.

Savers Unanticipated inflation hurts savers. As prices rise, the real value, or purchasing power, of an accumulation of savings deteriorates. Paper assets such as savings accounts, insurance policies, and annuities that were once adequate to meet rainy-day contingencies or provide for a comfortable retirement decline in real value during inflation. The simplest case is the person who hoards money as a cash balance. A $1,000 cash balance would have lost one-half its real value between 1985 and 2009. Of course, most forms of savings earn interest. But the value of savings will still decline if the rate of inflation exceeds the rate of interest.

Example: A household may save $1,000 in a certificate of deposit (CD) in a commercial bank or savings and loan association at 6 percent annual interest. But if inflation is 13 percent (as it was in 1980), the real value or purchasing power of that $1,000 will be cut to about $938 by the end of the year. Although the saver will receive $1,060 (equal to $1,000 plus $60 of interest), deflating that $1,060 for 13 percent inflation means that its real value is only about $938 (= $1,060 ÷ 1.13).

Creditors Unanticipated inflation harms creditors (lenders). Suppose Chase Bank lends Bob $1,000, to be repaid in 2 years. If in that time the price level doubles, the $1,000 that Bob repays will have only half the purchasing power of the $1,000 he borrowed. True, if we ignore interest charges, the same number of dollars will be repaid as was borrowed. But because of inflation, each of those dollars will buy only half as much as it did when the loan was negotiated. As prices go up, the purchasing power of the dollar goes down. So the borrower pays back less-valuable dollars than those received from the lender. The owners of Chase Bank suffer a loss of real income.

Who Is Unaffected or Helped by Inflation?

Some people are unaffected by inflation and others are actually helped by it. For the second group, inflation redistributes real income toward them and away from others.

Flexible-Income Receivers People who have flexible incomes may escape inflation's harm or even benefit from it. For example, individuals who derive their incomes solely from Social Security are largely unaffected by inflation because Social Security payments are *indexed* to the CPI. Benefits automatically increase when the CPI increases, preventing erosion of benefits from inflation. Some union workers also get automatic **cost-of-living adjustments (COLAs)** in their pay when the CPI rises, although such increases rarely equal the full percentage rise in inflation.

Some flexible-income receivers and all borrowers are helped by unanticipated inflation. The strong product demand and labor shortages implied by rapid demand-pull inflation may cause some nominal incomes to spurt ahead of the price level, thereby enhancing real incomes. For some, the 3 percent increase in nominal income that occurs when inflation is 2 percent may become a 7 percent increase when inflation is 5 percent. As an example, property owners faced with an inflation-induced real estate boom may be able to boost rents more rapidly than the rate of inflation. Also, some business owners may benefit from inflation. If product prices rise faster than resource prices, business revenues will increase more rapidly than costs. In those cases, the growth rate of profit incomes will outpace the rate of inflation.

Debtors Unanticipated inflation benefits debtors (borrowers). In our earlier example, Chase Bank's loss of real income from inflation is Bob's gain of real income. Debtor Bob borrows "dear" dollars but, because of inflation, pays back the principal and interest with "cheap" dollars whose purchasing power has been eroded by inflation. Real income is redistributed away from the owners of Chase Bank toward borrowers such as Bob.

The federal government, which had amassed $16.1 trillion of public debt through 2012, has also benefited from inflation. Historically, the federal government regularly paid off its loans by taking out new ones. Inflation permitted the Treasury to pay off its loans with dollars of less purchasing power than the dollars originally borrowed. Nominal national income and therefore tax collections rise with inflation; the amount of public debt owed does not. Thus, inflation reduces the real burden of the public debt to the federal government.

Anticipated Inflation

The redistribution effects of inflation are less severe or are eliminated altogether if people anticipate inflation and can adjust their nominal incomes to reflect the expected price-level rises. The prolonged inflation that began in the late 1960s prompted many labor unions in the 1970s to insist on labor contracts with cost-of-living adjustment clauses.

Similarly, if inflation is anticipated, the redistribution of income from lender to borrower may be altered. Suppose a lender (perhaps a commercial bank or a savings and loan institution) and a borrower (a household) both agree that 5 percent is a fair rate of interest on a 1-year loan provided the price level is stable. But assume that inflation has been occurring and is expected to be 6 percent over the next year. If the bank lends the household \$100 at 5 percent interest, the bank will be paid back \$105 at the end of the year. But if 6 percent inflation does occur during that year, the purchasing power of the \$105 will have been reduced to about \$99. The lender will, in effect, have paid the borrower \$1 for the use of the lender's money for a year.

The lender can avoid this subsidy by charging an *inflation premium*—that is, by raising the interest rate by 6 percent, the amount of the anticipated inflation. By charging 11 percent, the lender will receive back \$111 at the end of the year. Adjusted for the 6 percent inflation, that amount will have roughly the purchasing power of \$105 worth of today's money. The result then will be a mutually agreeable transfer of purchasing power from borrower to lender of \$5, or 5 percent, for the use of \$100 for 1 year. Financial institutions have also developed variable-interest-rate mortgages to protect themselves from the adverse effects of inflation. (Incidentally, this example points out that, rather than being a *cause* of inflation, high nominal interest rates are a *consequence* of inflation.)

Our example reveals the difference between the real rate of interest and the nominal rate of interest. The **real interest rate** is the percentage increase in *purchasing power* that the borrower pays the lender. In our example the real interest rate is 5 percent. The **nominal interest rate** is the percentage increase in *money* that the borrower pays the lender, including that resulting from the built-in expectation of inflation, if any. In equation form:

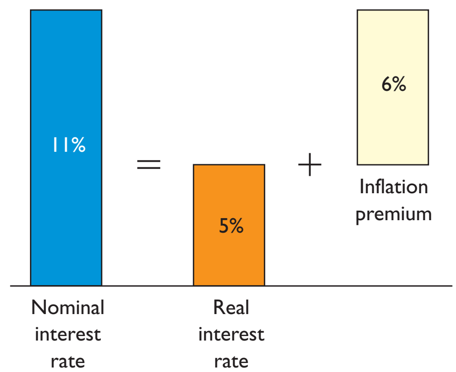
ORIGIN OF THE IDEA

O27.2

Real interest rates

Nominal interest rate = real interest rate + inflation premium (the expected rate of inflation)

FIGURE 27.5 The inflation premium and nominal and real interest rates. The inflation premium—the expected rate of inflation—gets built into the nominal interest rate. Here, the nominal interest rate of 11 percent comprises the real interest rate of 5 percent plus the inflation premium of 6 percent.



As illustrated in Figure 27.5, the nominal interest rate in our example is 11 percent.

Other Redistribution Issues

We end our discussion of the redistribution effects of inflation by making three final points:

- ***Deflation*** The effects of unanticipated deflation—declines in the price level—are the reverse of those of inflation. People with fixed nominal incomes will find their real incomes enhanced. Creditors will benefit at the expense of debtors. And savers will discover that the purchasing power of their savings has grown because of the falling prices.
- ***Mixed effects*** A person who is simultaneously an income earner, a holder of financial assets, and a debtor will probably find that the redistribution impact of unanticipated inflation is cushioned. If the person owns fixed-value monetary assets (savings accounts, bonds, and insurance policies), inflation will lessen their real value. But that same inflation may produce an increase in the person's nominal wage. Also, if the person holds a fixed-interest-rate mortgage, the real burden of that debt will decline. In short, many individuals are simultaneously hurt and helped by inflation. All these effects must be considered before we can conclude that any particular person's net position is better or worse because of inflation.
- ***Arbitrariness*** The redistribution effects of inflation occur regardless of society's goals and values. Inflation lacks a social conscience and takes from some and gives to others, whether they are rich, poor, young, old, healthy, or infirm.

QUICK REVIEW 27.4

- Inflation harms those who receive relatively fixed nominal incomes and either leaves unaffected or helps those who receive flexible nominal incomes.
- Unanticipated inflation hurts savers and creditors while benefiting debtors.
- The nominal interest rate equals the real interest rate plus the inflation premium (the expected rate of inflation).

Does Inflation Affect Output?

LO27.5 Discuss how inflation may affect the economy's level of real output.

Thus far, our discussion has focused on how inflation redistributes a specific level of total real income. But inflation also may affect an economy's level of real output (and thus its level of real income). The direction and significance of this effect on output depend on the type of inflation and its severity.

Cost-Push Inflation and Real Output

Recall that abrupt and unexpected rises in key resource prices such as oil can sufficiently drive up overall production costs to cause cost-push inflation. As prices rise, the quantity demanded of goods and services falls. So firms respond by producing less output, and unemployment goes up.

Economic events of the 1970s provide an example of how inflation can reduce real output. In late 1973 the Organization of Petroleum Exporting Countries (OPEC), by exerting its market power, managed to quadruple the price of oil. The cost-push inflationary effects generated rapid price-level increases in the 1973–1975 period. At the same time, the U.S. unemployment rate rose from slightly less than 5 percent in 1973 to 8.5 percent in 1975. Similar outcomes occurred in 1979–1980 in response to a second OPEC oil supply shock.

In short, cost-push inflation reduces real output. It redistributes a decreased level of real income.

Demand-Pull Inflation and Real Output

Economists do not fully agree on the effects of mild inflation (less than 3 percent) on real output. One perspective is that even low levels of inflation reduce real output because inflation diverts time and effort toward activities designed to hedge against inflation. Here are some examples of this:

- Businesses must incur the cost of changing thousands of prices on their shelves and in their computers simply to reflect inflation.
- Households and businesses must spend considerable time and effort obtaining the information they need to distinguish between real and nominal values such as prices, wages, and interest rates.
- To limit the loss of purchasing power from inflation, people try to limit the amount of money they hold in their billfolds and checking accounts at any one time and instead put more money into interest-bearing accounts and stock and bond funds. But cash and checks are needed in even greater amounts to buy the higher-priced goods and services. So more frequent trips, phone calls, or Internet visits to financial institutions are required to transfer funds to checking accounts and billfolds, when needed.

Without inflation, these uses of resources, time, and effort would not be needed, and they could be diverted toward producing more valuable goods and services. Proponents of "zero inflation" bolster their case by pointing to cross-country studies that indicate that lower rates of inflation are associated with higher rates of economic growth. Even mild inflation, say these economists, is detrimental to economic growth.

In contrast, other economists point out that full employment and economic growth depend on strong levels of total spending. Such spending creates high profits, strong demand for labor, and a powerful incentive for firms to expand their plants and equipment. In this view, the mild inflation that is a by-product of strong spending is a small price to pay for full employment and continued economic growth.

Moreover, a little inflation may have positive effects because it makes it easier for firms to adjust real wages downward when the demands for their products fall. With mild inflation, firms can reduce real wages by holding nominal wages steady. With zero inflation firms would need to cut nominal wages to reduce real wages. Such cuts in nominal wages are highly visible and may cause considerable worker resistance and labor strife.

Finally, defenders of mild inflation say that it is much better for an economy to err on the side of strong spending, full employment, economic growth, and mild inflation than on the side of weak spending, unemployment, recession, and deflation.

LAST WORD

Unemployment after the Great Recession

Economists Have Been Vigorously Debating Why Employment Recovered So Slowly after the Great Recession of 2007–2009.

The Great Recession began in December 2007 and ended in June 2009. The downturn was the most severe since the Great Depression of the 1930s, with real GDP falling 4.3 percent from peak to trough. After growth returned in mid-2009, real GDP increased slowly, taking until July 2011 to pass its prerecession peak.

Employment showed a similar pattern of rapid decline followed by slow recovery: 8.7 million people lost their jobs after employment peaked in January 2008. Employment began to expand again in early 2010, but job growth was so slow that in December 2012—more than three years after the recession ended and more than a year after real GDP had passed its prerecession high—employment was still 3.4 million less than it had been at the start of the recession.

The recession also dramatically increased the average length of time that workers spent unemployed before finding a new job. The typical (median) spell of unemployment went from lasting 7.7 weeks in June 2007 to a peak of 24.8 weeks in June 2010. By way of comparison, the highest previous measurement for this statistic had been 12.3 weeks during the 1981–1982 recession. Thus, the Great Recession saw not only the loss of 8.7 million jobs, but unprecedentedly long wait times for unemployed workers to find new jobs.

When real GDP initially fell by 4.3 percent, it was easy to understand why employers might have shed 8.7 million jobs: Fewer workers were needed to produce less output. But after real GDP recovered fully and passed its prerecession peak, economists began to debate why employment was still millions of jobs lower than it had been before the recession began.

Here are a few of the possible culprits.

Hyperinflation

All economists agree that **hyperinflation,** which is extraordinarily rapid inflation, can have a devastating impact on real output and employment.

As prices shoot up sharply and unevenly during hyperinflation, people begin to anticipate even more rapid inflation and normal economic relationships are disrupted. Business owners do not know what to charge for their products. Consumers do not know what to pay. Resource suppliers want to be paid with actual output, rather than with rapidly depreciating money. Money eventually becomes almost worthless and ceases to do its job as a medium of exchange. Businesses, anticipating further price increases, may find that hoarding both materials and finished products is profitable. Individual savers may decide to buy nonproductive wealth—jewels, gold and other precious metals, real estate, and so forth—rather than providing funds that can be borrowed to purchase capital equipment. The economy may be thrown into a state of barter, and production and exchange drop further. The net result is economic collapse and, often, political chaos.

Examples of hyperinflation are Germany after the First World War and Japan after the Second World War. In Germany, "prices increased so rapidly that waiters changed the prices on the menu several times during the course of a lunch. Sometimes customers had to pay double the price listed on the menu when they ordered."[2] In postwar Japan in 1947, "fisherman and farmers . . . used

[2]Theodore Morgan, *Income and Employment*, 2nd ed. (Englewood Cliffs, N.J.: Prentice Hall, 1952), p. 361.

Higher Federal Minimum Wage It is widely acknowledged that raising the minimum wage may increase unemployment by pricing low-productivity workers out of the labor market. Thus, one potential culprit for the slow recovery in employment was the July 2009 increase in the federal minimum wage from $6.55 to $7.25. But, at any given time, fewer than 3 percent of workers are employed at minimum-wage jobs. So it would be hard to blame the increase in the minimum wage for more than a very small fraction of the slow post-recession recovery in employment.

Longer Unemployment Benefits In November 2009, Congress decided to extend the maximum period of time that unemployed workers could draw unemployment benefits from 26 weeks to 99 weeks. That decision is believed by many economists to have affected a large enough fraction of unemployed workers to have contributed to the slow recovery in employment.

Congress had two intents when it extended the maximum draw period for unemployment benefits. The first was to help unemployed workers financially. The second was to help keep the economy moving by giving unemployed workers money to spend on goods and services.

One unintended consequence, however, was "inefficiently long search," meaning that many unemployed workers used the extended period during which they could survive on unemployment benefits to keep searching for perfect jobs even after they had been offered several so-so jobs. As a result, the unemployment rate stayed higher than it would have if benefits had continued to end after just 26 weeks and workers had felt financial pressure at an earlier date to accept so-so jobs rather than to keep on searching for perfect jobs.

Structural Adjustments Another explanation for the slow recovery in employment was that the economy required *structural adjustments*—changes in the basic structure of what was being produced and thus which industries needed workers. Consider the housing bubble that preceded the Great Recession. After the housing bubble collapsed, the economy needed to transition several million unemployed construction workers into other lines of work. Creating that many new jobs in other industries was going to take time. Thus, it was to be expected that employment was slow to recover after the recession ended.

Higher Labor Costs Other economists argued that worries about higher labor costs also contributed to the slow recovery in employment. In particular, they argued that several provisions of the 2010 health care reform law commonly known as Obamacare discouraged firms from hiring workers. One provision was an increase in the Medicare payroll tax. Another was the requirement that by 2014 any firm with more than 50 employees would have to provide health insurance coverage for all of its full-time workers.

That insurance provision was problematic because health insurance is very costly. In 2012, for example, the average cost for family coverage was $15,745 per worker. So, as the economy was making its way out of recession, it was the case that forward-looking employers may have reduced their hiring so as to have fewer full-time workers on the payroll when that provision of the law was scheduled to go into effect in 2014.

scales to weigh currency and change, rather than bothering to count it."[3]

There are also more recent examples: Between June 1986 and March 1991 the cumulative inflation in Nicaragua was 11,895,866,143 percent. From November 1993 to December 1994 the cumulative inflation rate in the Democratic Republic of Congo was 69,502 percent. From February 1993 to January 1994 the cumulative inflation rate in Serbia was 156,312,790 percent.[4]

Such dramatic hyperinflations are always the consequence of highly imprudent expansions of the money supply by government. The rocketing money supply produces frenzied total spending and severe demand-pull inflation. Zimbabwe's 14.9 billion percent inflation in 2008 is just the latest example.

QUICK REVIEW 27.5

- Cost-push inflation reduces real output and employment.
- Economists argue about the effects of demand-pull inflation. Some argue that even mild demand-pull inflation (1 to 3 percent) reduces the economy's real output. Other say that mild inflation may be a necessary by-product of the high and growing spending that produces high levels of output, full employment, and economic growth.
- Hyperinflation, caused by highly imprudent expansions of the money supply, may undermine the monetary system and cause severe declines in real output.

[3]Raburn M. Williams, *Inflation! Money, Jobs, and Politicians* (Arlington Heights, Ill.: AHM Publishing, 1980), p. 2.

[4]Stanley Fischer, Ratna Sahay, and Carlos Végh, "Modern Hyper- and High Inflations," *Journal of Economic Literature*, September 2002, p. 840.

SUMMARY

LO27.1 Describe the business cycle and its primary phases.

The United States and other industrial economies have gone through periods of fluctuations in real GDP, employment, and the price level. Although they have certain phases in common—peak, recession, trough, expansion—business cycles vary greatly in duration and intensity.

Although economists explain the business cycle in terms of underlying causal factors such as major innovations, productivity shocks, money creation, and financial crises, they generally agree that changes in the level of total spending are the immediate causes of fluctuating real output and employment.

The business cycle affects all sectors of the economy, though in varying ways and degrees. The cycle has greater effects on output and employment in the capital goods and durable consumer goods industries than in the services and nondurable goods industries.

LO27.2 Illustrate how unemployment is measured and explain the different types of unemployment.

Economists distinguish between frictional, structural, and cyclical unemployment. The full-employment or natural rate of unemployment, which is made up of frictional and structural unemployment, is currently between 5 and 6 percent. The presence of part-time and discouraged workers makes it difficult to measure unemployment accurately.

The GDP gap, which can be either a positive or a negative value, is found by subtracting potential GDP from actual GDP. The economic cost of unemployment, as measured by the GDP gap, consists of the goods and services forgone by society when its resources are involuntarily idle. Okun's law suggests that every 1-percentage-point increase in unemployment above the natural rate causes an additional 2 percent negative GDP gap.

LO27.3 Explain how inflation is measured and distinguish between cost-push inflation and demand-pull inflation.

Inflation is a rise in the general price level and is measured in the United States by the Consumer Price Index (CPI). When inflation occurs, each dollar of income will buy fewer goods and services than before. That is, inflation reduces the purchasing power of money. Deflation is a decline in the general price level.

Unemployment rates and inflation rates vary widely globally. Unemployment rates differ because nations have different natural rates of unemployment and often are in different phases of their business cycles. Inflation and unemployment rates in the United States recently have been in the middle to low range compared with rates in other industrial nations.

Economists discern both demand-pull and cost-push (supply-side) inflation. Demand-pull inflation results from an excess of total spending relative to the economy's capacity to produce. The main source of cost-push inflation is abrupt and rapid increases in the prices of key resources. These supply shocks push up per-unit production costs and ultimately raise the prices of consumer goods.

LO27.4 Relate how unanticipated inflation can redistribute real income.

Unanticipated inflation arbitrarily redistributes real income at the expense of fixed-income receivers, creditors, and savers. If inflation is anticipated, individuals and businesses may be able to take steps to lessen or eliminate adverse redistribution effects.

When inflation is anticipated, lenders add an inflation premium to the interest rate charged on loans. The nominal interest rate thus reflects the real interest rate plus the inflation premium (the expected rate of inflation).

LO27.5 Discuss how inflation may affect the economy's level of real output.

Cost-push inflation reduces real output and employment. Proponents of zero inflation argue that even mild demand-pull inflation (1 to 3 percent) reduces the economy's real output. Other economists say that mild inflation may be a necessary byproduct of the high and growing spending that produces high levels of output, full employment, and economic growth.

Hyperinflation, caused by highly imprudent expansions of the money supply, may undermine the monetary system and cause severe declines in real output.

TERMS AND CONCEPTS

business cycles
peak
recession
trough
expansion
labor force
unemployment rate
discouraged workers
frictional unemployment
structural unemployment
cyclical unemployment
full-employment rate of unemployment
natural rate of unemployment (NRU)
potential output
GDP gap
Okun's law
inflation
Consumer Price Index (CPI)
deflation
demand-pull inflation
cost-push inflation
per-unit production costs
core inflation
nominal income

real income
unanticipated inflation
anticipated inflation
cost-of-living adjustments (COLAs)
real interest rate
nominal interest rate
hyperinflation

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What are the four phases of the business cycle? How long do business cycles last? Why does the business cycle affect output and employment in capital goods industries and consumer durable goods industries more severely than in industries producing consumer nondurables? **LO27.1**
2. How, in general, can a financial crisis lead to a recession? How, in general, can a major new invention lead to an expansion? **LO27.1**
3. How is the labor force defined and who measures it? How is the unemployment rate calculated? Does an increase in the unemployment rate necessarily mean a decline in the size of the labor force? Why is a positive unemployment rate—one more than zero percent—fully compatible with full employment? **LO27.2**
4. How, in general, do unemployment rates vary by race and ethnicity, gender, occupation, and education? Why does the average length of time people are unemployed rise during a recession? **LO27.2**
5. Why is it difficult to distinguish between frictional, structural, and cyclical unemployment? Why is unemployment an economic problem? What are the consequences of a negative GDP gap? What are the noneconomic effects of unemployment? **LO27.2**
6. Because the United States has an unemployment compensation program that provides income for those out of work, why should we worry about unemployment? **LO27.2**
7. What is the Consumer Price Index (CPI) and how is it determined each month? How does the Bureau of Labor Statistics calculate the rate of inflation from one year to the next? What effect does inflation have on the purchasing power of a dollar? How does it explain differences between nominal and real interest rates? How does deflation differ from inflation? **LO27.3**
8. Distinguish between demand-pull inflation and cost-push inflation. Which of the two types is most likely to be associated with a negative GDP gap? Which with a positive GDP gap, in which actual GDP exceeds potential GDP? What is core inflation? Why is it calculated? **LO27.3**
9. Explain how an increase in your nominal income and a decrease in your real income might occur simultaneously. Who loses from inflation? Who gains? **LO27.4**
10. Explain how hyperinflation might lead to a severe decline in total output. **LO27.5**
11. **LAST WORD** Why was the 2009 hike in the minimum wage probably not responsible for much of the slow growth in employment after the Great Recession? What is inefficiently long search and how is it affected by the duration of unemployment benefits? How might Obamacare have discouraged hiring?

REVIEW QUESTIONS

1. Place the phases of the business cycle in order. **LO27.1**
   Recession
   Trough
   Peak
   Expansion
2. Most economists agree that the immediate cause of the large majority of cyclical changes in the levels of real output and employment is unexpected changes in __________________. **LO27.1**
   a. The level of total spending.
   b. The level of the stock market.
   c. The level of the trade deficit.
   d. The level of unemployment.
3. Suppose that an economy has 9 million people working full-time. It also has 1 million people who are actively seeking work but currently unemployed as well as 2 million discouraged workers who have given up looking for work and are currently unemployed. What is this economy's unemployment rate? **LO27.2**
   a. 10 percent.
   b. 15 percent.
   c. 20 percent.
   d. 25 percent.
4. Label each of the following scenarios as either frictional unemployment, structural unemployment, or cyclical unemployment. **LO27.2**
   a. Tim just graduated and is looking for a job.
   b. A recession causes a local factory to lay off 30 workers.
   c. Thousands of bus and truck drivers permanently lose their jobs when driverless, computer-driven vehicles make human drivers redundant.
   d. Hundreds of New York legal jobs permanently disappear when a lot of legal work gets outsourced to lawyers in India.

5. The unemployment rate that is consistent with full employment is known as ______________. **LO27.2**
   a. The natural rate of unemployment.
   b. The unnatural rate of unemployment.
   c. The status quo rate of unemployment.
   d. Cyclical unemployment.
   e. Okun's rate of unemployment.
6. A country's current unemployment rate is 11 percent. Economists estimate that its natural rate of unemployment is 6 percent. About how large is this economy's negative GDP gap? **LO27.2**
   a. 1 percent.
   b. 3 percent.
   c. 6 percent.
   d. 10 percent.
7. Cost-push inflation occurs when there is ______________. **LO27.3**
   a. Excess inventory.
   b. A trade deficit.
   c. Rising per-unit production costs.
   d. Excess demand for goods and services.
8. Jimmer's nominal income will go up by 10 percent next year. Inflation is expected to be −2 percent next year. By approximately how much will Jimmer's real income change next year? **LO27.3**
   a. −2 percent.
   b. 8 percent.
   c. 10 percent.
   d. 12 percent.
9. Kaitlin has $10,000 of savings that she may deposit with her local bank. Kaitlin wants to earn a real rate of return of at least 4 percent and she is expecting inflation to be exactly 3 percent. What is the lowest nominal interest rate that Kaitlin would be willing to accept from her local bank? **LO27.4**
   a. 4 percent.
   b. 5 percent.
   c. 6 percent.
   d. 7 percent.
10. True or False: Lenders are helped by unanticipated inflation. **LO27.4**
11. Economists agree that ______________ inflation reduces real output. **LO27.5**
   a. Cost-push.
   b. Demand-pull.
   c. Push-pull.

PROBLEMS

1. Suppose that a country's annual growth rates were 5, 3, 4, −1, −2, 2, 3, 4, 6, and 3 in yearly sequence over a 10-year period. What was the country's trend rate of growth over this period? Which set of years most clearly demonstrates an expansionary phase of the business cycle? Which set of years best illustrates a recessionary phase of the business cycle? **LO27.1**
2. Assume the following data for a country: total population, 500; population under 16 years of age or institutionalized, 120; not in labor force, 150; unemployed, 23; part-time workers looking for full-time jobs, 10. What is the size of the labor force? What is the official unemployment rate? **LO27.2**
3. Suppose that the natural rate of unemployment in a particular year is 5 percent and the actual rate of unemployment is 9 percent. Use Okun's law to determine the size of the GDP gap in percentage-point terms. If the potential GDP is $500 billion in that year, how much output is being forgone because of cyclical unemployment? **LO27.2**
4. If the CPI was 110 last year and is 121 this year, what is this year's rate of inflation? In contrast, suppose that the CPI was 110 last year and is 108 this year. What is this year's rate of inflation? What term do economists use to describe this second outcome? **LO27.3**
5. How long would it take for the price level to double if inflation persisted at (*a*) 2 percent per year, (*b*) 5 percent per year, and (*c*) 10 percent per year? **LO27.3**
6. If your nominal income rose by 5.3 percent and the price level rose by 3.8 percent in some year, by what percentage would your real income (approximately) increase? If your nominal income rose by 2.8 percent and your real income rose by 1.1 percent in some year, what must have been the (approximate) rate of inflation? **LO27.4**
7. Suppose that the nominal rate of inflation is 4 percent and the inflation premium is 2 percent. What is the real interest rate? Alternatively, assume that the real interest rate is 1 percent and the nominal interest rate is 6 percent. What is the inflation premium? **LO27.4**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **EIGHT**

MACROECONOMIC MODELS AND FISCAL POLICY

CHAPTER **28**

Basic Macroeconomic Relationships*

Learning Objectives

LO28.1 Describe how changes in income affect consumption (and saving).

LO28.2 List and explain factors other than income that can affect consumption.

LO28.3 Explain how changes in real interest rates affect investment.

LO28.4 Identify and explain factors other than the real interest rate that can affect investment.

LO28.5 Illustrate how changes in investment (or one of the other components of total spending) can increase or decrease real GDP by a multiple amount.

In Chapter 27 we discussed the business cycle, unemployment, and inflation. Our eventual goal is to build economic models that can explain these phenomena. This chapter begins that process by examining the basic relationships that exist between three different pairs of economic aggregates. (Recall that to economists "aggregate" means "total" or "combined.") Specifically, this chapter looks at the relationships between:

- income and consumption (and income and saving).
- the interest rate and investment.
- changes in spending and changes in output.

*Note to the Instructor: If you wish to bypass the aggregate expenditures model (Keynesian cross model) covered in full in Chapter 29, assigning the present chapter will provide a seamless transition to the AD-AS model of Chapter 30 and the chapters beyond. If you want to cover the aggregate expenditures model, this present chapter provides the necessary building blocks.

What explains the trends in consumption (consumer spending) and saving reported in the news? How do changes in interest rates affect investment? How can initial changes in spending ultimately produce multiplied changes in GDP? The basic macroeconomic relationships discussed in this chapter answer these questions.

The Income-Consumption and Income-Saving Relationships

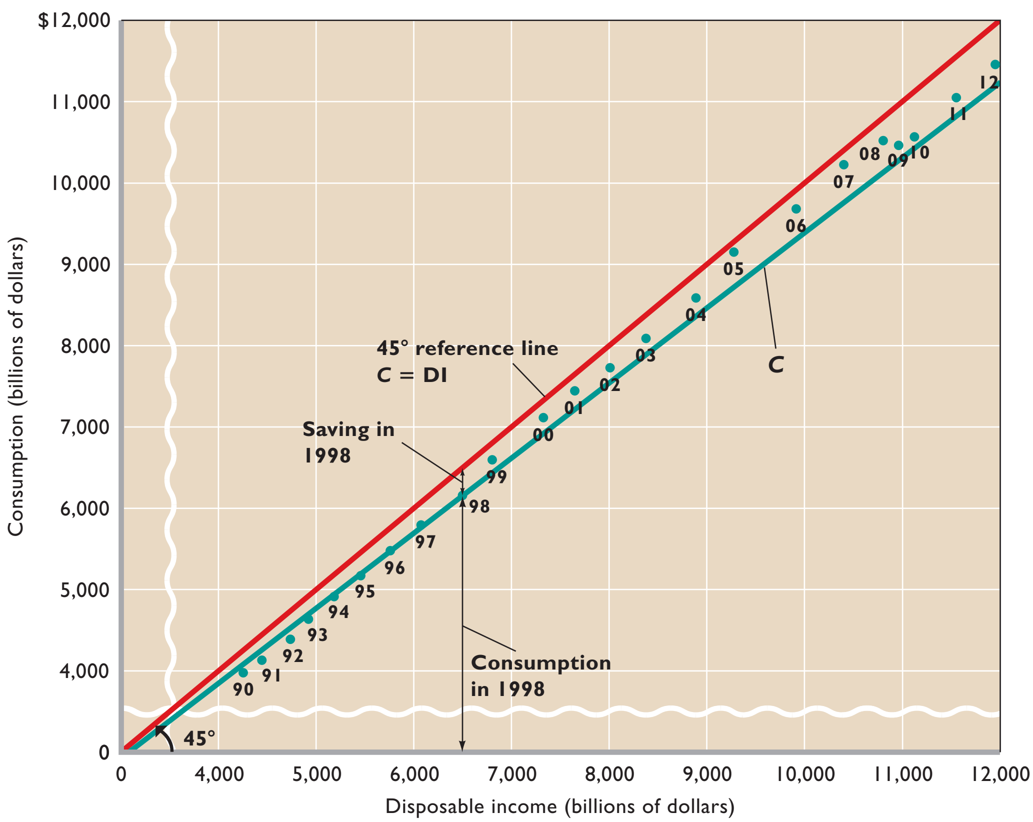
LO28.1 Describe how changes in income affect consumption (and saving).

The other-things-equal relationship between income and consumption is one of the best-established relationships in macroeconomics. In examining that relationship, we are also exploring the relationship between income and saving. Recall that economists define *personal saving* as "not spending" or as "that part of disposable (after-tax) income not consumed." Saving (*S*) equals disposable income (DI) *minus* consumption (*C*).

Many factors determine a nation's levels of consumption and saving, but the most significant is disposable income. Consider some recent historical data for the United States. In Figure 28.1 each dot represents consumption and disposable income for 1 year since 1987. The line *C* that is loosely fitted to these points shows that consumption is directly (positively) related to disposable income; moreover, households spend most of their income.

But we can say more. The **45°(degree) line** is a reference line. Because it bisects the 90° angle formed by the two axes of the graph, each point on it is equidistant from the two axes. At each point on the 45° line, consumption would equal disposable income, or $C =$ DI. Therefore, the vertical

FIGURE 28.1 Consumption and disposable income, 1990–2012. Each dot in this figure shows consumption and disposable income in a specific year. The line *C*, which generalizes the relationship between consumption and disposable income, indicates a direct relationship and shows that households consume most of their after-tax incomes.



Source: Bureau of Economic Analysis, **www.bea.gov**.

distance between the 45° line and any point on the horizontal axis measures either consumption *or* disposable income. If we let it measure disposable income, the vertical distance between it and the consumption line labeled *C* represents the amount of saving (*S*) in that year. Saving is the amount by which actual consumption in any year falls short of the 45° line—($S = \text{DI} - C$). For example, in 1998 disposable income was \$6,498.9 billion and consumption was \$6,157.5 billion, so saving was \$341.4 billion. Observe that the vertical distance between the 45° line and line *C* increases as we move rightward along the horizontal axis and decreases as we move leftward. Like consumption, saving typically varies directly with the level of disposable income. However, that historical pattern broke down somewhat in several years preceding the recession of 2007–2009.

The Consumption Schedule

The dots in Figure 28.1 represent historical data—the actual amounts of DI, *C*, and *S* in the United States over a period of years. But, because we want to understand how the economy would behave under different possible scenarios, we need a schedule showing the various amounts that households would *plan* to consume at each of the various levels of disposable income that might prevail at some specific time. Columns 1 and 2 of Table 28.1, represented in **Figure 28.2a (Key Graph),** show the hypothetical consumption schedule that we require. This **consumption schedule** (or "consumption function") reflects the direct consumption–disposable income relationship suggested by the data in Figure 28.1, and it is consistent with many household budget studies. In the aggregate, households increase their spending as their disposable income rises and spend a larger proportion of a small disposable income than of a large disposable income.

ORIGIN OF THE IDEA

O28.1

Income-consumption relationship

The Saving Schedule

It is relatively easy to derive a **saving schedule** (or "saving function"). Because saving equals disposable income less consumption ($S = \text{DI} - C$), we need only subtract consumption (Table 28.1, column 2) from disposable income (column 1) to find the amount saved (column 3) at each DI. Thus, columns 1 and 3 in Table 28.1 are the saving schedule, represented in Figure 28.2b. The graph shows that there is a direct relationship between saving and DI but that saving is a smaller proportion of a small DI than of a large DI. If households consume a smaller and smaller proportion of DI as DI increases, then they must be saving a larger and larger proportion.

Remembering that at each point on the 45° line consumption equals DI, we see that *dissaving* (consuming in excess of after-tax income) will occur at relatively low DIs. For example, at \$370 billion (row 1, Table 28.1), consumption is \$375 billion. Households can consume more than their current incomes by liquidating (selling for cash) accumulated wealth or by borrowing. Graphically, dissaving is shown as the vertical distance of the consumption schedule above the 45° line or as the vertical distance of the saving schedule

TABLE 28.1 **Consumption and Saving Schedules (in Billions) and Propensities to Consume and Save**

(1) Level of Output and Income (GDP = DI)	(2) Consumption (*C*)	(3) Saving (*S*), (1) − (2)	(4) Average Propensity to Consume (APC), (2)/(1)	(5) Average Propensity to Save (APS), (3)/(1)	(6) Marginal Propensity to Consume (MPC), Δ(2)/Δ(1)*	(7) Marginal Propensity to Save (MPS), Δ(3)/Δ(1)*
(1) \$370	\$375	\$−5	1.01	−.01	.75	.25
(2) 390	390	0	1.00	.00	.75	.25
(3) 410	405	5	.99	.01	.75	.25
(4) 430	420	10	.98	.02	.75	.25
(5) 450	435	15	.97	.03	.75	.25
(6) 470	450	20	.96	.04	.75	.25
(7) 490	465	25	.95	.05	.75	.25
(8) 510	480	30	.94	.06	.75	.25
(9) 530	495	35	.93	.07	.75	.25
(10) 550	510	40	.93	.07		

*The Greek letter Δ, delta, means "the change in."

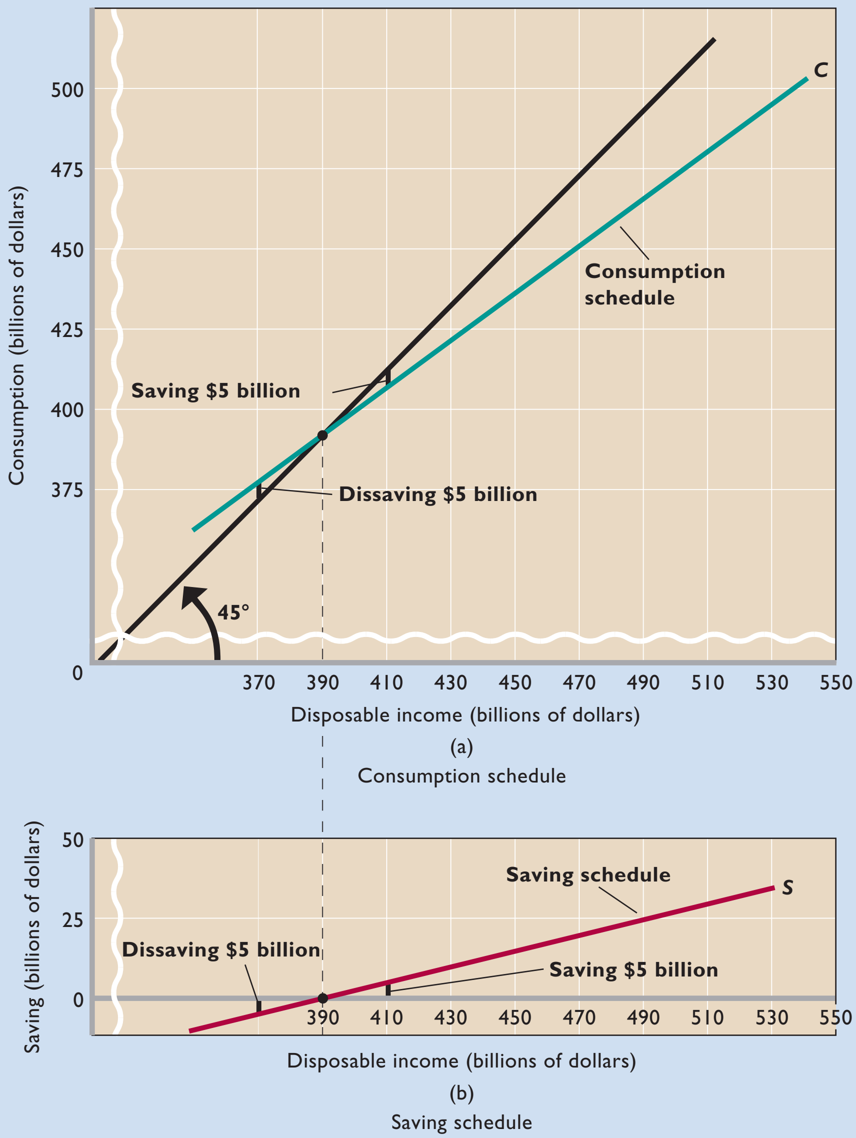


FIGURE 28.2 Consumption and saving schedules. The two parts of this figure show the income-consumption and income-saving relationships in Table 28.1 graphically. (a) Consumption rises as income increases. Saving is negative (dissaving occurs) when the consumption schedule is above the 45° line, and saving is positive when the consumption schedule is below the 45° line. (b) Like consumption, saving increases as income goes up. The saving schedule is found by subtracting the consumption schedule in the top graph vertically from the 45° line. For these hypothetical data, saving is −$5 billion at $370 billion of income, zero at $390 billion of income, and $5 billion at $410 billion of income. Saving is zero where consumption equals disposable income.

QUICK QUIZ FOR FIGURE 28.2

1. The slope of the consumption schedule in this figure is 0.75. Thus, the:
   a. slope of the saving schedule is 1.33.
   b. marginal propensity to consume is 0.75.
   c. average propensity to consume is 0.25.
   d. slope of the saving schedule is also 0.75.
2. In this figure, when consumption is a positive amount, saving:
   a. must be a negative amount.
   b. must also be a positive amount.
   c. can be either a positive or a negative amount.
   d. is zero.
3. In this figure:
   a. the marginal propensity to consume is constant at all levels of income.
   b. the marginal propensity to save rises as disposable income rises.
   c. consumption is inversely (negatively) related to disposable income.
   d. saving is inversely (negatively) related to disposable income.
4. When consumption equals disposable income:
   a. the marginal propensity to consume is zero.
   b. the average propensity to consume is zero.
   c. consumption and saving must be equal.
   d. saving must be zero.

Answers: 1. b; 2. c; 3. a; 4. d

below the horizontal axis. We have marked the dissaving at the $370 billion level of income in Figure 28.2a and 28.2b. Both vertical distances measure the $5 billion of dissaving that occurs at $370 billion of income.

In our example, the **break-even income** is $390 billion (row 2, Table 28.1). This is the income level at which households plan to consume their entire incomes (C = DI). Graphically, the consumption schedule cuts the 45° line, and the saving schedule cuts the horizontal axis (saving is zero) at the break-even income level.

At all higher incomes, households plan to save part of their incomes. Graphically, the vertical distance between the consumption schedule and the 45° line measures this saving (see Figure 28.2a), as does the vertical distance between the saving schedule and the horizontal axis (see Figure 28.2b). For example, at the $410 billion level of income (row 3, Table 28.1), both these distances indicate $5 billion of saving.

Average and Marginal Propensities

Columns 4 to 7 in Table 28.1 show additional characteristics of the consumption and saving schedules.

APC and APS The fraction, or percentage, of total income that is consumed is the **average propensity to consume (APC).** The fraction of total income that is saved is the **average propensity to save (APS).** That is,

$$\text{APC} = \frac{\text{consumption}}{\text{income}}$$

and

$$\text{APS} = \frac{\text{saving}}{\text{income}}$$

For example, at $470 billion of income (row 6, Table 28.1), the APC is $\frac{450}{470} = \frac{45}{47}$, or about 0.96 (= 96 percent), while the APS is $\frac{20}{470} = \frac{2}{47}$, or about 0.04 (= 4 percent). Columns 4 and 5 in Table 28.1 show the APC and APS at each of the 10 levels of DI. As implied by our previous discussion, the APC falls as DI increases, while the APS rises as DI goes up.

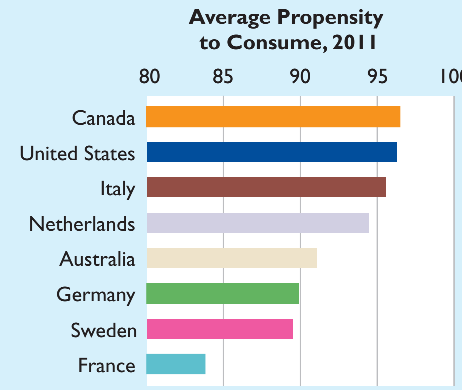
Because disposable income is either consumed or saved, the fraction of any DI consumed plus the fraction saved (not consumed) must exhaust that income. Mathematically, APC + APS = 1 at any level of disposable income, as columns 4 and 5 in Table 28.1 illustrate. So if 0.96 of the $470 billion of income in row 6 in consumed, 0.04 must be saved. That is why APC + APS = 1.

Global Perspective 28.1 shows APCs for several countries.

GLOBAL PERSPECTIVE 28.1

Average Propensities to Consume, Selected Nations

There are surprisingly large differences in average propensities to consume (APCs) among nations. In 2011, Italy, the United States, Canada, and the Netherlands in particular had substantially higher APCs, and thus lower APSs, than several other advanced economies.



Source: Organization for Economic Cooperation and Development, OECD, **www.oecd.org**. Derived from OECD household saving rates as percentages of disposable income. Econ Outlook 86, Annex Table 23, extracted March 2013.

MPC and MPS The fact that households consume a certain proportion of a particular total income, for example, $\frac{45}{47}$ of a $470 billion disposable income, does not guarantee they will consume the same proportion of any *change* in income they might receive. The proportion, or fraction, of any change in income consumed is called the **marginal propensity to consume (MPC),** "marginal" meaning "extra" or "a change in." Equivalently, the MPC is the ratio of a change in consumption to a change in the income that caused the consumption change:

$$\text{MPC} = \frac{\text{change in consumption}}{\text{change in income}}$$

Similarly, the fraction of any change in income saved is the **marginal propensity to save (MPS).** The MPS is the ratio of a change in saving to the change in income that brought it about:

$$\text{MPS} = \frac{\text{change in saving}}{\text{change in income}}$$

If disposable income is $470 billion (row 6 horizontally in Table 28.1) and household income rises by $20 billion to $490 billion (row 7), households will consume $\frac{15}{20}$, or $\frac{3}{4}$, and save $\frac{5}{20}$, or $\frac{1}{4}$, of that increase in income. In other words, the MPC is $\frac{3}{4}$ or 0.75, and the MPS is $\frac{1}{4}$ or 0.25, as shown in columns 6 and 7.

The sum of the MPC and the MPS for any change in disposable income must always be 1. Consuming or saving out of extra income is an either-or proposition; the fraction of any change in income not consumed is, by definition, saved. If 0.75 of extra disposable income is consumed, 0.25 must be saved. The fraction consumed (MPC) plus the fraction saved (MPS) must exhaust the whole change in income:

$$\text{MPC} + \text{MPS} = 1$$

In our example, 0.75 plus 0.25 equals 1.

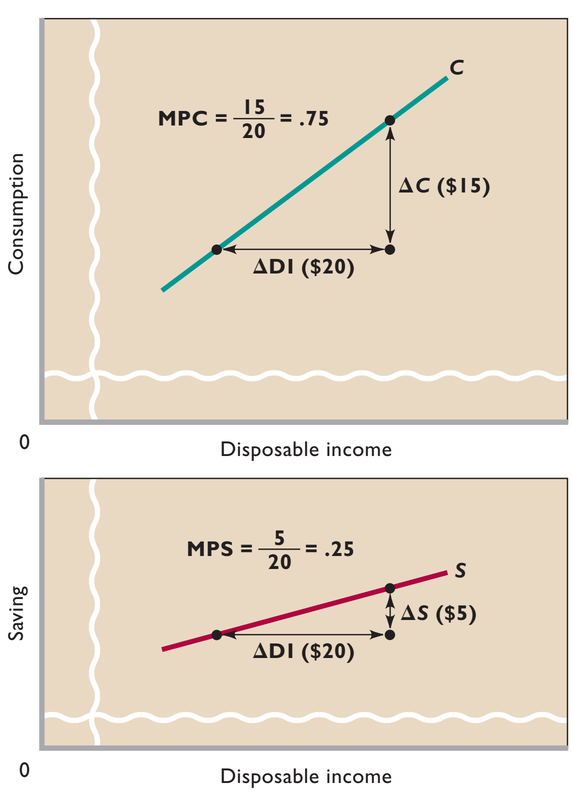
MPC and MPS as Slopes The MPC is the numerical value of the slope of the consumption schedule, and the MPS is the numerical value of the slope of the saving schedule. We know from the appendix to Chapter 1 that the slope of any line is the ratio of the vertical change to the horizontal change occasioned in moving from one point to another on that line.

WORKED PROBLEMS

W28.1
Consumption and saving

Figure 28.3 measures the slopes of the consumption and saving lines, using enlarged portions of Figure 28.2a and 28.2b. Observe that consumption changes by $15 billion (the vertical change) for each $20 billion change in disposable income (the horizontal change). The slope of the consumption line is thus 0.75 (= $15/$20), which is the value of the MPC. Saving changes by $5 billion (shown as the vertical change) for every $20 billion change in disposable income (shown as the horizontal change). The slope of the saving line therefore is 0.25 (= $5/$20), which is the value of the MPS.

FIGURE 28.3 The marginal propensity to consume and the marginal propensity to save. In the two parts of this figure, the Greek letter delta (Δ) means "the change in." (a) The MPC is the slope (Δ*C*/ΔDI) of the consumption schedule. (b) The MPS is the slope (Δ*S*/ΔDI) of the saving schedule.



Nonincome Determinants of Consumption and Saving

LO28.2 List and explain factors other than income that can affect consumption.

The amount of disposable income is the basic determinant of the amounts households will consume and save. But certain determinants other than income might prompt households to consume more or less at each possible level of income and thereby change the locations of the consumption and saving schedules. Those other determinants are wealth, borrowing, expectations, and interest rates.

- ***Wealth*** A household's wealth is the dollar amount of all the assets that it owns minus the dollar amount of its liabilities (all the debt that it owes). Households build wealth by saving money out of current income. The point of building wealth is to increase consumption possibilities. The larger the stock of wealth that a household can build up, the larger will be its present and future consumption possibilities.

  Events sometimes suddenly boost the value of existing wealth. When this happens, households tend to increase their spending and reduce their saving. This so-called **wealth effect** shifts the consumption schedule upward and the saving schedule downward. They move in response to households taking advantage of the increased consumption possibilities afforded by the sudden increase in wealth. Examples: In the late 1990s, skyrocketing U.S. stock values expanded the value of household wealth by increasing the value of household assets. Predictably, households spent more and saved less. In contrast, a strong "reverse wealth effect" occurred in 2008. Plunging real estate and stock market prices joined together to erase $11.2 trillion (yes, trillion) of household wealth. Consumers quickly reacted by reducing their consumption spending. The consumption schedule shifted downward.
- ***Borrowing*** Household borrowing also affects consumption. When a household borrows, it can

increase current consumption beyond what would be possible if its spending were limited to its disposable income. By allowing households to spend more, borrowing shifts the current consumption schedule upward.

But note that there is no "free lunch." While borrowing in the present allows for higher consumption in the present, it necessitates lower consumption in the future when the debts that are incurred due to the borrowing must be repaid. Stated a bit differently, increased borrowing increases debt (liabilities), which in turn reduces household wealth (since *wealth* = *assets* − *liabilities*). This reduction in wealth reduces future consumption possibilities in much the same way that a decline in asset values would. But note that the term "reverse wealth effect" is reserved for situations in which wealth unexpectedly changes because asset values unexpectedly change. It is not used to refer to situations such as the one being discussed here where wealth is intentionally reduced by households through borrowing and piling up debt to increase current consumption.

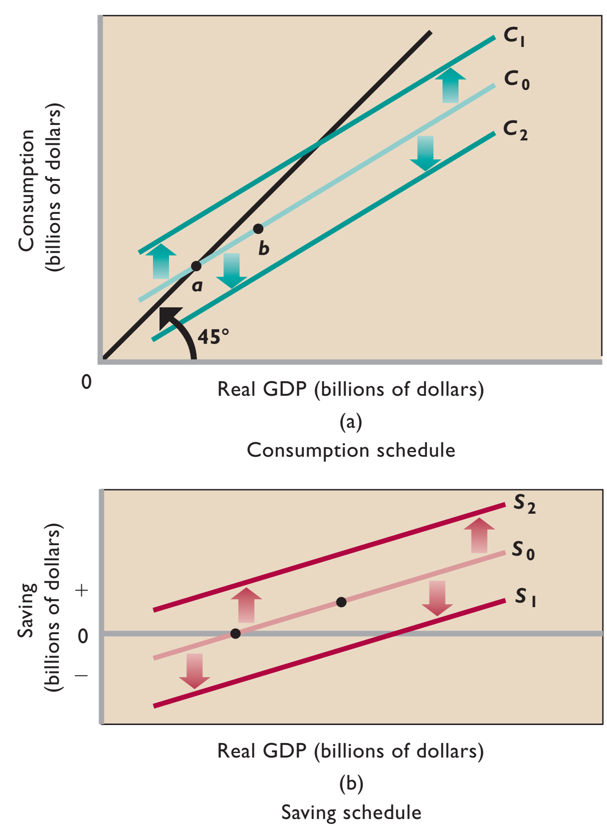
- ***Expectations*** Household expectations about future prices and income may affect current spending and saving. For example, the expectation of higher prices tomorrow may cause households to buy more today while prices are still low. Thus, the current consumption schedule shifts up and the current saving schedule shifts down. Or expectations of a recession and thus lower income in the future may lead households to reduce consumption and save more today. Their greater present saving will help build wealth that will help them ride out the expected bad times. The consumption schedule therefore will shift down and the saving schedule will shift up.
- ***Real interest rates*** When real interest rates (those adjusted for inflation) fall, households tend to borrow more, consume more, and save less. A lower interest rate, for example, decreases monthly loan payments and induces consumers to purchase automobiles and other goods bought on credit. A lower interest rate also diminishes the incentive to save because of the reduced interest "payment" to the saver. These effects on consumption and saving, however, are very modest. They mainly shift consumption toward some products (those bought on credit) and away from others. At best, lower interest rates shift the consumption schedule slightly upward and the saving schedule slightly downward. Higher interest rates do the opposite.

Other Important Considerations

There are several additional important points regarding the consumption and saving schedules:

- ***Switching to real GDP*** When developing macroeconomic models, economists change their focus from the relationship between consumption (and saving) and *disposable income* to the relationship between consumption (and saving) and *real domestic output (real GDP)*. This modification is reflected in Figure 28.4a and 28.4b, where the horizontal axes measure real GDP.

FIGURE 28.4 Shifts of the (a) consumption and (b) saving schedules. Normally, if households consume more at each level of real GDP, they are necessarily saving less. Graphically this means that an upward shift of the consumption schedule (C_0 to C_1) entails a downward shift of the saving schedule (S_0 to S_1). If households consume less at each level of real GDP, they are saving more. A downward shift of the consumption schedule (C_0 to C_2) is reflected in an upward shift of the saving schedule (S_0 to S_2). This pattern breaks down, however, when taxes change; then the consumption and saving schedules move in the *same* direction—opposite to the direction of the tax change.



CONSIDER THIS . . .

The Great Recession and the Paradox of Thrift

The Great Recession of 2007–2009 altered the prior consumption and saving behavior in the economy. Concerned about reduced wealth, high debt, and potential job losses, households increased their saving and reduced their consumption at each level of after-tax income (or each level of GDP). In Figure 28.4, this outcome is illustrated as the downward *shift* of the consumption schedule in the top graph and the upward *shift* of the saving schedule in the lower graph.

This change of behavior illustrates the so-called **paradox of thrift,** which refers to the possibility that a recession can be made worse when households become more thrifty and save in response to the downturn. The paradox of thrift rests on two major ironies. One irony is that saving more is *good* for the economy in the long run, as noted in Chapter 1 and Chapter 24. It finances investment and therefore fuels subsequent economic growth. But saving more can be *bad* for the economy during a recession. Because firms are pessimistic about future sales, the increased saving is not likely to be matched by an equal amount of added investment. The extra saving simply reduces spending on currently produced goods and services. That means that even more businesses suffer, more layoffs occur, and people's incomes decline even more.

The paradox of thrift has a second irony related to the *fallacy of composition* (Chapter 1, Last Word): Households as a group may inadvertently end up saving less when each individual household tries to save more during a recession. This is because each household's attempt to save more implies that it is also attempting to spend less. Across all households, that collective reduction in total spending in the economy creates additional job losses and further drives down total income. The decline in total income reduces the ability of households as a group to save as much as they did before their spending reduction and subsequent income decline.

- ***Changes along schedules*** The movement from one point to another on a consumption schedule (for example, from *a* to *b* on C_0 in Figure 28.4a) is a *change in the amount consumed* and is solely caused by a change in real GDP. On the other hand, an upward or downward shift of the entire schedule, for example, a shift from C_0 to C_1 or C_2 in Figure 28.4a, is a *shift of the consumption schedule* and is caused by changes in any one or more of the *nonincome* determinants of consumption just discussed.

  A similar distinction in terminology applies to the saving schedule in Figure 28.4b.
- ***Simultaneous shifts*** Changes in wealth, expectations, interest rates, and household debt will shift the consumption schedule in one direction and the saving schedule in the opposite direction. If households decide to consume more at each possible level of real GDP, they must save less, and vice versa. (Even when they spend more by borrowing, they are, in effect, reducing their current saving by the amount borrowed since borrowing is, effectively, "negative saving.") Graphically, if the consumption schedule shifts upward from C_0 to C_1 in Figure 28.4a, the saving schedule shifts downward, from S_0 to S_1 in Figure 28.4b. Similarly, a downward shift of the consumption schedule from C_0 to C_2 means an upward shift of the saving schedule from S_0 to S_2.
- ***Taxation*** In contrast, a change in taxes shifts the consumption and saving schedules in the same direction. Taxes are paid partly at the expense of consumption and partly at the expense of saving. So an increase in taxes will reduce both consumption and saving, shifting the consumption schedule in Figure 28.4a and the saving schedule in Figure 28.4b downward. Conversely, households will partly consume and partly save any decrease in taxes. Both the consumption schedule and saving schedule will shift upward.
- ***Stability*** The consumption and saving schedules usually are relatively stable unless altered by major tax increases or decreases. Their stability may be because consumption-saving decisions are strongly influenced by long-term considerations such as saving to meet emergencies or saving for retirement. It may also be because changes in the nonincome determinants frequently work in opposite directions and therefore may be self-canceling.

QUICK REVIEW 28.1

- Both consumption spending and saving rise when disposable income increases; both fall when disposable income decreases.
- The average propensity to consume (APC) is the fraction of any specific level of disposable income that is spent on consumer goods; the average propensity to save (APS) is the fraction of any specific level of disposable income that is saved. The APC falls and the APS rises as disposable income increases.
- The marginal propensity to consume (MPC) is the fraction of a change in disposable income that is consumed and it is the slope of the consumption schedule; the marginal propensity to save (MPS) is the fraction of a change in disposable income that is saved and it is the slope of the saving schedule.
- Changes in consumer wealth, consumer expectations, interest rates, household debt, and taxes can shift the consumption and saving schedules (as they relate to real GDP).

The Interest-Rate–Investment Relationship

LO28.3 Explain how changes in real interest rates affect investment.

In our consideration of major macro relationships, we next turn to the relationship between the real interest rate and investment. Recall that investment consists of expenditures on new plants, capital equipment, machinery, inventories, and so on. The investment decision is a marginal-benefit–marginal-cost decision: The marginal benefit from investment is the expected rate of return businesses hope to realize. The marginal cost is the interest rate that must be paid for borrowed funds. Businesses will invest in all projects for which the expected rate of return exceeds the interest rate. Expected returns (profits) and the interest rate therefore are the two basic determinants of investment spending.

Expected Rate of Return

Investment spending is guided by the profit motive; businesses buy capital goods only when they think such purchases will be profitable. Suppose the owner of a small cabinetmaking shop is considering whether to invest in a new sanding machine that costs $1,000 and has a useful life of only 1 year. (Extending the life of the machine beyond 1 year complicates the economic decision but does not change the fundamental analysis. We discuss the valuation of returns beyond 1 year in Chapter 35.) The new machine will increase the firm's output and sales revenue. Suppose the net expected revenue from the machine (that is, after such operating costs as power, lumber, labor, and certain taxes have been subtracted) is $1,100. Then, after the $1,000 cost of the machine is subtracted from the net expected revenue of $1,100, the firm will have an expected profit of $100. Dividing this $100 profit by the $1,000 cost of the machine, we find that the **expected rate of return,** r, on the machine is 10 percent (= $100/$1,000). It is important to note that this is an *expected* rate of return, not a *guaranteed* rate of return. The investment may or may not generate as much revenue or as much profit as anticipated. Investment involves risk.

The Real Interest Rate

One important cost associated with investing that our example has ignored is interest, which is the financial cost of borrowing the $1,000 of *money* "capital" to purchase the $1,000 of *real* capital (the sanding machine).

The interest cost of the investment is computed by multiplying the interest rate, i, by the $1,000 borrowed to buy the machine. If the interest rate is, say, 7 percent, the total interest cost will be $70. This compares favorably with the net expected return of $100, which produced the 10 percent expected rate of return. If the investment works out as expected, it will add $30 to the firm's profit.

We can generalize as follows: If the expected rate of return (10 percent) exceeds the interest rate (here, 7 percent), the investment should be undertaken. The firm expects the investment to be profitable. But if the interest rate (say, 12 percent) exceeds the expected rate of return (10 percent), the investment should not be undertaken. The firm expects the investment to be unprofitable. The firm should undertake all investment projects it thinks will be profitable. This means that the firm should array its prospective investment projects from the highest expected rate of return, r, downward and then invest in all projects for which r exceeds i. The firm therefore should invest to the point where $r = i$ because then it will have undertaken all investments for which r is greater than i.

This guideline applies even if a firm finances the investment internally out of funds saved from past profit rather than borrowing the funds. The role of the interest rate in the investment decision does not change. When the firm uses money from savings to invest in the sander, it incurs an opportunity cost because it forgoes the interest income it could have earned by lending the funds to someone else. That interest cost, converted to percentage terms, needs to be weighed against the expected rate of return.

The *real* rate of interest, rather than the *nominal* rate, is crucial in making investment decisions. Recall from Chapter 27 that the nominal interest rate is expressed in dollars of current value, while the real interest rate is stated in dollars of constant or inflation-adjusted value. Recall that the real interest rate is the nominal rate less the rate of inflation. In our sanding machine illustration, our implicit assumption of a constant price level ensures that all our data, including the interest rate, are in real terms.

But what if inflation *is* occurring? Suppose a $1,000 investment is expected to yield a real (inflation-adjusted) rate of return of 10 percent and the nominal interest rate is 15 percent. At first, we would say the investment would be unprofitable. But assume there is ongoing inflation of 10 percent per year. This means the investing firm will pay back dollars with approximately 10 percent less in purchasing power. While the nominal interest rate is 15 percent, the real rate is only 5 percent (= 15 percent − 10 percent). By comparing this 5 percent real interest rate with the 10 percent expected real rate of return, we find that the investment is potentially profitable and should be undertaken.

Investment Demand Curve

We now move from a single firm's investment decision to total demand for investment goods by the entire business sector. Assume that every firm has estimated the expected rates of return from all investment projects and has recorded those data. We can cumulate (successively sum) these data by asking: How many dollars' worth of investment projects have an expected rate of return of, say, 16 percent or more? How many have 14 percent or more? How many have 12 percent or more? And so on.

Suppose no prospective investments yield an expected return of 16 percent or more. But suppose there are $5 billion of investment opportunities with expected rates of return between 14 and 16 percent; an additional $5 billion yielding between 12 and 14 percent; still an additional $5 billion yielding between 10 and 12 percent; and an additional $5 billion in each successive 2 percent range of yield down to and including the 0 to 2 percent range.

To cumulate these figures for each rate of return, r, we add the amounts of investment that will yield each particular rate of return r or higher. This provides the data in the table in Figure 28.5. The data are shown graphically in **Figure 28.5 (Key Graph).** In the table, the number opposite 12 percent, for example, means there are $10 billion of investment opportunities that will yield an expected rate of return of 12 percent or more. The $10 billion includes the $5 billion of investment expected to yield a return of 14 percent or more plus the $5 billion expected to yield between 12 and 14 percent.

We know from our example of the sanding machine that an investment project will be undertaken if its expected rate of return, r, exceeds the real interest rate, i. Let's first suppose i is 12 percent. Businesses will undertake all investments for which r exceeds 12 percent. That is, they will invest until the 12 percent rate of return equals the 12 percent interest rate. Figure 28.5 reveals that $10 billion of investment spending will be undertaken at a 12 percent interest rate; that means $10 billion of investment projects have an expected rate of return of 12 percent or more.

Put another way: At a financial "price" of 12 percent, $10 billion of investment goods will be demanded. If the interest rate is lower, say, 8 percent, the amount of investment for which r equals or exceeds i is $20 billion. Thus, firms will demand $20 billion of investment goods at an 8 percent real interest rate. At 6 percent, they will demand $25 billion of investment goods.

By applying the marginal-benefit–marginal-cost rule that investment projects should be undertaken up to the point where $r = i$, we see that we can add the real interest rate to the vertical axis in Figure 28.5. The curve in Figure 28.5 not only shows rates of return; it shows the quantity of investment demanded at each "price" i (interest rate) of investment. The vertical axis in Figure 28.5 shows the various possible real interest rates, and the horizontal axis shows the corresponding quantities of investment demanded. The inverse (downsloping) relationship between the interest rate (price) and dollar quantity of investment demanded conforms to the law of demand discussed in Chapter 3. The curve *ID* in Figure 28.5 is the economy's **investment demand curve.** It shows the amount of investment forthcoming at each real interest rate. The level of investment depends on the expected rate of return and the real interest rate.

ORIGIN OF THE IDEA

O28.2

Interest-rate–investment relationship

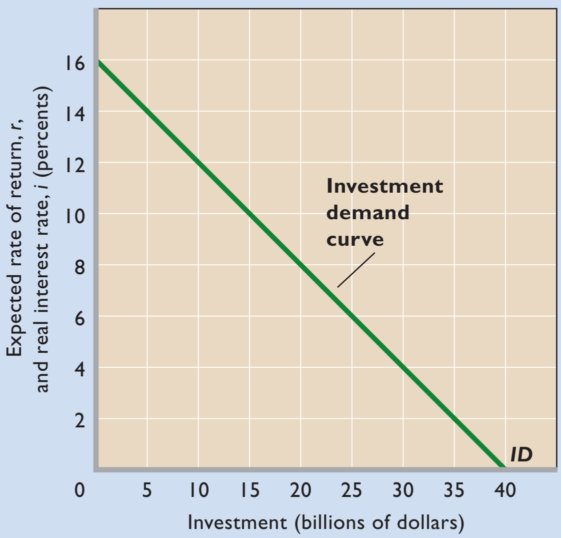
Shifts of the Investment Demand Curve

LO28.4 Identify and explain factors other than the real interest rate that can affect investment.

Figure 28.5 shows the relationship between the interest rate and the amount of investment demanded, other things equal. When other things change, the investment

KEY GRAPH

FIGURE 28.5 The investment demand curve. The investment demand curve is constructed by arraying all potential investment projects in descending order of their expected rates of return. The curve slopes downward, reflecting an inverse relationship between the real interest rate (the financial "price" of each dollar of investing) and the quantity of investment demanded.



Real Interest Rate (*i*) and Expected Rate of Return (*r*)	Cumulative Amount of Investment Having This Rate of Return or Higher, Billions per Year
16%	$ 0
14	5
12	10
10	15
8	20
6	25
4	30
2	35
0	40

QUICK QUIZ FOR FIGURE 28.5

1. The investment demand curve:
- **a.** reflects a direct (positive) relationship between the real interest rate and investment.
- **b.** reflects an inverse (negative) relationship between the real interest rate and investment.
- **c.** shifts to the right when the real interest rate rises.
- **d.** shifts to the left when the real interest rate rises.

2. In this figure:
- **a.** greater cumulative amounts of investment are associated with lower real interest rates.
- **b.** lesser cumulative amounts of investment are associated with lower expected rates of return on investment.
- **c.** higher interest rates are associated with higher expected rates of return on investment, and therefore greater amounts of investment.
- **d.** interest rates and investment move in the same direction.

3. In this figure, if the real interest rate falls from 6 to 4 percent:
- **a.** investment will increase from 0 to $30 billion.
- **b.** investment will decrease by $5 billion.
- **c.** the expected rate of return will rise by $5 billion.
- **d.** investment will increase from $25 billion to $30 billion.

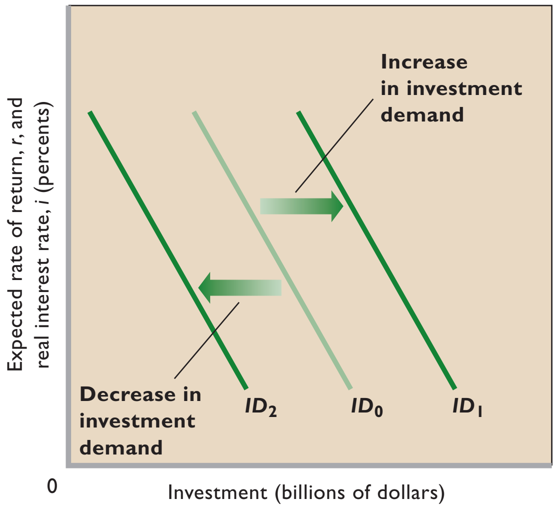
4. In this figure, investment will be:
- **a.** zero if the real interest rate is zero.
- **b.** $40 billion if the real interest rate is 16 percent.
- **c.** $30 billion if the real interest rate is 4 percent.
- **d.** $20 billion if the real interest rate is 12 percent.

Answers: 1. b; 2. a; 3. d; 4. c

demand curve shifts. In general, any factor that leads businesses collectively to expect greater rates of return on their investments increases investment demand. That factor shifts the investment demand curve to the right, as from ID_0 to ID_1 in Figure 28.6. Any factor that leads businesses collectively to expect lower rates of return on their investments shifts the curve to the left, as from ID_0 to ID_2. What are those non-interest-rate determinants of investment demand?

- ***Acquisition, maintenance, and operating costs*** The initial costs of capital goods, and the estimated costs of operating and maintaining those goods, affect the expected rate of return on investment. When these costs rise, the expected rate of return from prospective investment projects falls and the investment demand curve shifts to the left. Example: Higher electricity costs associated with operating tools and machinery shifts the investment

FIGURE 28.6 Shifts of the investment demand curve. Increases in investment demand are shown as rightward shifts of the investment demand curve; decreases in investment demand are shown as leftward shifts of the investment demand curve.



demand curve to the left. Lower costs, in contrast, shift it to the right.

- ***Business taxes*** When government is considered, firms look to expected returns *after taxes* in making their investment decisions. An increase in business taxes lowers the expected profitability of investments and shifts the investment demand curve to the left; a reduction of business taxes shifts it to the right.
- ***Technological change*** Technological progress—the development of new products, improvements in existing products, and the creation of new machinery and production processes—stimulates investment. The development of a more efficient machine, for example, lowers production costs or improves product quality and increases the expected rate of return from investing in the machine. Profitable new products (cholesterol medications, Internet services, high-definition televisions, cellular phones, and so on) induce a flurry of investment as businesses tool up for expanded production. A rapid rate of technological progress shifts the investment demand curve to the right.
- ***Stock of capital goods on hand*** The stock of capital goods on hand, relative to output and sales, influences investment decisions by firms. When the economy is overstocked with production facilities and when firms have excessive inventories of finished goods, the expected rate of return on new investment declines. Firms with excess production capacity have little incentive to invest in new capital. Therefore, less investment is forthcoming at each real interest rate; the investment demand curve shifts leftward.

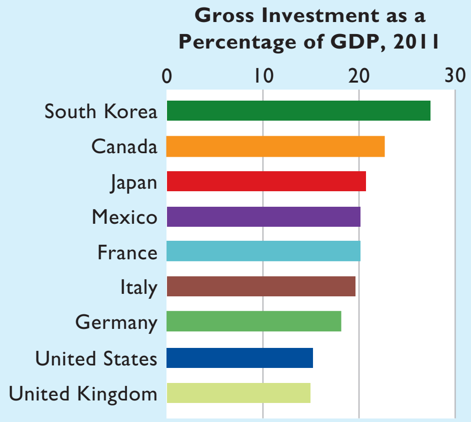
  When the economy is understocked with production facilities and when firms are selling their output as fast as they can produce it, the expected rate of return on new investment increases and the investment demand curve shifts rightward.
- ***Planned inventory changes*** Recall from Chapter 25 that the definition of investment includes changes in inventories of unsold goods. An increase in inventories is counted as positive investment while a decrease in inventories is counted as negative investment. It is important to remember that some inventory changes are planned, while others are unplanned. Since the investment demand curve deals only with *planned* investment, it is only affected by *planned* changes that firms desire to make to their inventory levels. If firms are planning to increase their inventories, the investment demand curve shifts to the right. If firms are planning on decreasing their inventories, the investment demand curve shifts to the left.

  Firms make planned changes to their inventory levels mostly because they are expecting either faster or slower sales. A firm that expects its sales to double in the next year will want to keep more inventory in stock, thereby increasing its investment demand. By contrast, a firm that is expecting slower sales will plan on reducing its inventory, thereby reducing its overall investment demand. But because life often does not turn out as expected, firms often find that the actual amount of inventory investment that they end up making is either more or less than what they had planned. The size of the gap is, naturally, the dollar amount of their *unplanned* inventory changes. These unplanned inventory adjustments will play a large role in the aggregate expenditures model studied in Chapter 29.
- ***Expectations*** We noted that business investment is based on expected returns (expected additions to profit). Most capital goods are durable, with a life expectancy of 10 or 20 years. Thus, the expected rate of return on capital investment depends on the firm's expectations of future sales, future operating costs, and future profitability of the product that the capital helps produce. These expectations are based on forecasts of future business conditions as well as on such elusive and difficult-to-predict factors as changes in

GLOBAL PERSPECTIVE 28.2

Gross Investment Expenditures as a Percentage of GDP, Selected Nations

As a percentage of GDP, investment varies widely by nation. These differences, of course, can change from year to year.



Source: Gross fixed capital formation data from International Financial Statistics, International Monetary Fund, **www.imf.org**.

the domestic political climate, international relations, population growth, and consumer tastes. If executives become more optimistic about future sales, costs, and profits, the investment demand curve will shift to the right; a pessimistic outlook will shift the curve to the left.

Global Perspective 28.2 compares investment spending relative to GDP for several nations in a recent year. Domestic real interest rates and investment demand determine the levels of investment relative to GDP.

Instability of Investment

In contrast to consumption, investment is unstable; it rises and falls quite often. Investment, in fact, is the most volatile component of total spending—so much so that most of the fluctuations in output and employment that happen over the course of the business cycle can be attributed to demand shocks relating to unexpected increases and decreases in investment. Figure 28.7 shows just how volatile investment in the United States has been. Notice that the percentage swings in investment in real terms are greater than the percentage swings in real GDP. Several interrelated factors drawn from our previous discussion explain the variability of investment.

- ***Variability of expectations*** Business expectations can change quickly when some event suggests a significant possible change in future business conditions. Changes in exchange rates, trade barriers, legislative actions, stock market prices, government economic policies, the outlook for war or peace, court decisions in key labor or antitrust cases, and a host of similar considerations may cause substantial shifts in business expectations.
- ***Durability*** Because of their durability, capital goods have indefinite useful lifespans. Within limits, purchases of capital goods are discretionary and therefore can be postponed. Firms can scrap or replace older equipment and buildings, or they can patch them up and use them for a few more years. Optimism about the future may prompt firms to replace

CONSIDER THIS . . .

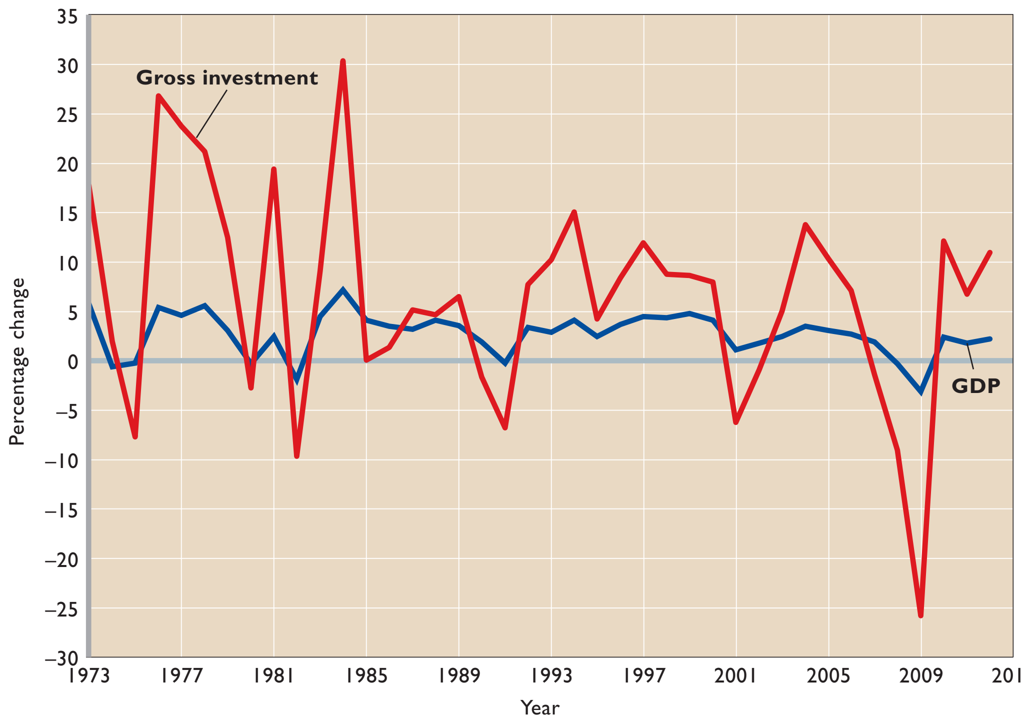
The Great Recession and the Investment Riddle

During the severe recession of 2007–2009, real interest rates essentially declined to zero. Figure 28.5 suggests that this drop in interest rates should have boosted investment spending. But investment declined substantially during this period. On an annual basis, it declined by 9 percent in 2008 and 26 percent in 2009. Does this combination of lower real interest rates and reduced investment make Figure 28.5 irrelevant?

Definitely not! The key to the investment riddle is that during the recession, the investment demand curve shifted inward, as from ID_0 to ID_2 in Figure 28.6. This inward shift overwhelmed any investment-increasing effects of the decline of real interest rates. The net result turned out to be less investment, not more.

The leftward shift of the investment demand reflected a decline in the expected returns from investment. Firms envisioned zero or negative returns on investment in new capital because they were facing an overstock of existing capital relative to their current sales. Understandably, they therefore were not inclined to invest. Also, firms were extremely pessimistic about when the economy would regain its strength. This pessimism also contributed to low expected rates of return on investment and thus to exceptionally weak investment demand. Further, even though the interest rate was so low, firms that wanted to borrow and invest found that lenders were reluctant to lend them money for fear that they would not be able to pay back the loans.

FIGURE 28.7 The volatility of investment, 1973–2012. Annual percentage changes in investment spending are often several times greater than the percentage changes in GDP. (Data are in real terms. Investment is gross private domestic investment).



Source: Bureau of Economic Analysis, **www.bea.gov**.

their older facilities and such modernizing will call for a high level of investment. A less optimistic view, however, may lead to smaller amounts of investment as firms repair older facilities and keep them in use.

- ***Irregularity of innovation*** New products and processes stimulate investment. Major innovations such as railroads, electricity, airplanes, automobiles, computers, the Internet, and cell phones induce vast upsurges or "waves" of investment spending that in time recede. But such innovations occur quite irregularly, adding to the volatility of investment.
- ***Variability of profits*** High current profits often generate optimism about the future profitability of new investments, whereas low current profits or losses spawn considerable doubt about the wisdom of new investments. Additionally, firms often save a portion of current profits as retained earnings and use these funds (as well as borrowed funds) to finance new investments. So current profits affect both the incentive and ability to invest. But profits themselves are highly variable from year to year, contributing to the volatility of investment.

In terms of our previous analysis, we would represent volatility of investment as occasional and substantial unexpected shifts of the investment demand curve (as in Figure 28.6), which cause significant changes in investment spending (as in Figure 28.7). These demand shocks can contribute to cyclical instability.

QUICK REVIEW 28.2

- A specific investment will be undertaken if the expected rate of return, r, equals or exceeds the real interest rate, i.
- The investment demand curve shows the total monetary amounts that will be invested by an economy at various possible real interest rates.
- The investment demand curve shifts when changes occur in (a) the costs of acquiring, operating, and maintaining capital goods, (b) business taxes, (c) technology, (d) the stock of capital goods on hand, and (e) business expectations.

The Multiplier Effect*

LO28.5 Illustrate how changes in investment (or one of the other components of total spending) can increase or decrease real GDP by a multiple amount.

A final basic relationship that requires discussion is the relationship between changes in spending and changes in real GDP. Assuming that the economy has room to expand—so that increases in spending do not lead to increases in prices—there is a direct relationship between these two aggregates. More spending results in a higher GDP; less spending results in a lower GDP. But there is much more to this relationship. A change in spending, say, investment, ultimately changes output and income by more than the initial change in investment spending. That surprising result is called the *multiplier effect:* a change in a component of total spending leads to a larger change in GDP. The **multiplier** determines how much larger that change will be; it is the ratio of a change in GDP to the initial change in spending (in this case, investment). Stated generally,

$$\text{Multiplier} = \frac{\text{change in real GDP}}{\text{initial change in spending}}$$

By rearranging this equation, we can also say that

$$\text{Change in GDP} = \text{multiplier} \times \text{initial change in spending}$$

So if investment in an economy rises by \$30 billion and GDP increases by \$90 billion as a result, we then know from our first equation that the multiplier is 3 (= \$90/\$30).

Note these three points about the multiplier:

- The "initial change in spending" is usually associated with investment spending because of investment's volatility. But changes in consumption (unrelated to changes in income), net exports, and government purchases also lead to the multiplier effect.
- The "initial change in spending" associated with investment spending results from a change in the real interest rate and/or a shift of the investment demand curve.
- Implicit in the preceding point is that the multiplier works in both directions. An increase in initial spending will create a multiple increase in GDP, while a decrease in spending will create a multiple decrease in GDP.

*Instructors who cover the full aggregate expenditures (AE) model (Chapter 29) rather than moving directly to aggregate demand and aggregate supply (Chapter 30) may choose to defer this discussion until after the analysis of equilibrium real GDP.

Rationale

The multiplier effect follows from two facts. First, the economy supports repetitive, continuous flows of expenditures and income through which dollars spent by Smith are received as income by Chin and then spent by Chin and received as income by Gonzales, and so on. (This chapter's Last Word presents this idea in a humorous way.) Second, any change in income will change both consumption and saving in the same direction as, and by a fraction of, the change in income.

It follows that an initial change in spending will set off a spending chain throughout the economy. That chain of spending, although of diminishing importance at each successive step, will cumulate to a multiple change in GDP. Initial changes in spending produce magnified changes in output and income.

The table in Figure 28.8 illustrates the rationale underlying the multiplier effect. Suppose that a \$5 billion increase in investment spending occurs. We assume that the MPC is 0.75, the MPS is 0.25, and prices remain constant. That is, neither the initial increase in spending nor any of the subsequent increases in spending will cause prices to rise.

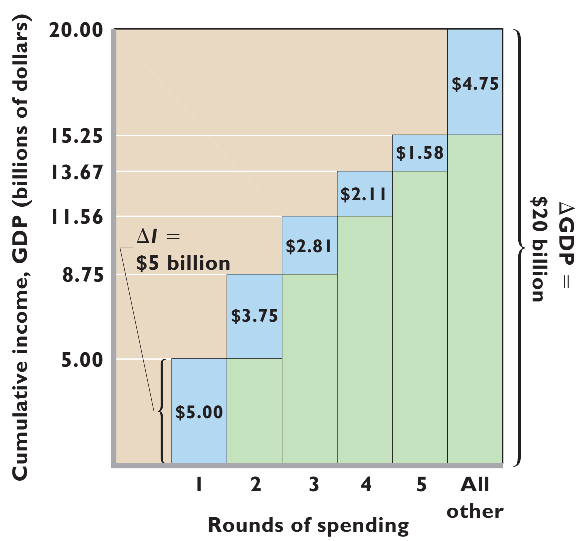
The initial \$5 billion increase in investment generates an equal amount of wage, rent, interest, and profit income because spending and receiving income are two sides of the same transaction. How much consumption will be induced by this \$5 billion increase in the incomes of households? We find the answer by applying the marginal propensity to consume of 0.75 to this change in income. Thus, the \$5 billion increase in income initially raises consumption by \$3.75 (= 0.75 × \$5) billion and saving by \$1.25 (= 0.25 × \$5) billion, as shown in columns 2 and 3 in the table.

Other households receive as income (second round) the \$3.75 billion of consumption spending. Those households consume 0.75 of this \$3.75 billion, or \$2.81 billion, and save 0.25 of it, or \$0.94 billion. The \$2.81 billion that is consumed flows to still other households as income to be spent or saved (third round). And the process continues, with the added consumption and income becoming less in each round. The process ends when there is no more additional income to spend.

The bar chart in Figure 28.8 shows several rounds of the multiplier process of the table graphically. As shown by rounds 1 to 5, each round adds a smaller and smaller blue block to national income and GDP. The process, of course, continues beyond the five rounds shown (for convenience we have simply cumulated the subsequent declining blocks into a single block labeled "All other"). The

FIGURE 28.8 The multiplier process (MPC = 0.75). An initial change in investment spending of \$5 billion creates an equal \$5 billion of new income in round 1. Households spend \$3.75 (= 0.75 × \$5) billion of this new income, creating \$3.75 billion of added income in round 2. Of this \$3.75 billion of new income, households spend \$2.81 (= 0.75 × \$3.75) billion, and income rises by that amount in round 3. Such income increments over the entire process get successively smaller but eventually produce a total change of income and GDP of \$20 billion. The multiplier therefore is 4 (= \$20 billion/\$5 billion).

	(1) Change in Income	(2) Change in Consumption (MPC = 0.75)	(3) Change in Saving (MPS = 0.25)
Increase in investment of **\$5.00**	\$5.00	\$ 3.75	\$1.25
Second round	3.75	2.81	0.94
Third round	2.81	2.11	0.70
Fourth round	2.11	1.58	0.53
Fifth round	1.58	1.19	0.39
All other rounds	4.75	3.56	1.19
Total	**\$20.00**	\$15.00	\$5.00



accumulation of the additional income in each round—the sum of the blue blocks—is the total change in income or GDP resulting from the initial \$5 billion change in spending. Because the spending and respending effects of the increase in investment diminish with each successive round of spending, the cumulative increase in output and income eventually ends. In this case, the ending occurs when \$20 billion of additional income accumulates. Thus, the multiplier is 4 (= \$20 billion/\$5 billion).

The Multiplier and the Marginal Propensities

You may have sensed from the table in Figure 28.8 that the fractions of an increase in income consumed (MPC) and saved (MPS) determine the cumulative respending effects of any initial change in spending and therefore determine the size of the multiplier. The MPC and the multiplier are directly related and the MPS and the multiplier are inversely related. The precise formulas are as shown in the next two equations:

$$\text{Multiplier} = \frac{1}{1 - \text{MPC}}$$

Recall, too, that MPC + MPS = 1. Therefore MPS = 1 − MPC, which means we can also write the multiplier formula as

$$\text{Multiplier} = \frac{1}{\text{MPS}}$$

This latter formula is a quick way to determine the multiplier. All you need to know is the MPS.

The smaller the fraction of any change in income saved, the greater the respending at each round and, therefore, the greater the multiplier. When the MPS is 0.25, as in our example, the multiplier is 4. If the MPS were 0.2, the multiplier would be 5. If the MPS were 0.33, the multiplier would be 3. Let's see why.

Suppose the MPS is 0.2 and businesses increase investment by \$5 billion. In the first round of the table in Figure 28.8, consumption will rise by \$4 billion (= MPC of 0.8 × \$5 billion) rather than by \$3.75 billion because saving will increase by \$1 billion (= MPS of 0.2 × \$5 billion) rather than \$1.25 billion. The greater rise in consumption in round 1 will produce a greater increase in income in round 2. The same will be true for all successive rounds. If we worked through all rounds of the multiplier, we would find that the process ends when income has cumulatively increased by \$25 billion, not the \$20 billion shown in the table. When the MPS is 0.2 rather than 0.25, the multiplier is 5 (= \$25 billion/\$5 billion) as opposed to 4 (= \$20 billion/\$5 billion.)

If the MPS were 0.33 rather than 0.25, the successive increases in consumption and income would be less than those in the table in Figure 28.8. We would discover that the process ended with a \$15 billion increase in income rather than the \$20 billion shown. When the MPS is 0.33, the multiplier is 3 (= \$15 billion/\$5 billion). The mathematics works such that the multiplier is equal to the reciprocal of the MPS. The reciprocal of any number is the quotient you obtain by dividing 1 by that number.

A large MPC (small MPS) means the succeeding rounds of consumption spending shown in Figure 28.8 diminish slowly and thereby cumulate to a large change in income. Conversely, a small MPC (a large MPS) causes the increases in consumption to decline quickly, so the cumulative change in income is small. The relationship between the MPC (and thus the MPS) and the multiplier is summarized in Figure 28.9.

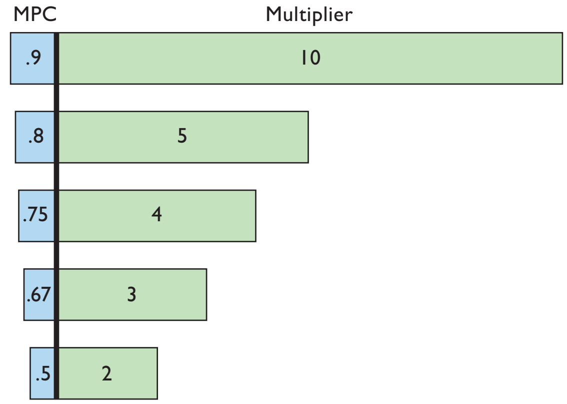
WORKED PROBLEMS

W28.2
Multiplier effect

How Large Is the Actual Multiplier Effect?

The multiplier we have just described is based on simplifying assumptions. Consumption of domestic output rises by the increases in income minus the increases in saving. But in reality, consumption of domestic output increases in each round by a lesser amount than implied by the MPS alone. In addition to saving, households use some of the extra income in each round to purchase additional goods from abroad (imports) and pay additional taxes. Buying imports and paying taxes drains off some of the additional consumption spending (on domestic output) created by the increases in income. So the multiplier effect is reduced and the 1/MPS formula for the multiplier overstates the actual outcome. To correct that problem, we would need to change the multiplier equation to read "1 divided by the fraction of the change in income that is not spent on domestic output." Also, we will find in later chapters that an increase in spending may be partly dissipated as inflation rather than realized fully as an increase in real GDP. This happens when increases in spending drive up prices. The multiplier process still happens, but it induces a much smaller change in real output because, at higher prices, any given amount of spending buys less real output. Economists disagree on the size of the actual multiplier in the United States. Estimates range from as high as 2.5 to as low as zero. So keep in mind throughout later discussions that the actual multiplier is much lower than the multipliers in our simple explanatory examples.

FIGURE 28.9 The MPC and the multiplier. The larger the MPC (the smaller the MPS), the greater the size of the multiplier.



QUICK REVIEW 28.3

- The multiplier effect reveals that an initial change in spending can cause a larger change in domestic income and output. The multiplier is the factor by which the initial change is magnified: multiplier = change in real GDP/initial change in spending.
- The higher the marginal propensity to consume (the lower the marginal propensity to save), the larger the multiplier: multiplier = 1/(1 − MPC) or 1/MPS.
- Economists disagree on the size of the actual multiplier in the U.S. economy; estimates range all the way from zero to 2.5.

LAST WORD

Squaring the Economic Circle

Humorist Art Buchwald Examines the Multiplier

WASHINGTON—The recession hit so fast that nobody knows exactly how it happened. One day we were the land of milk and honey and the next day we were the land of sour cream and food stamps.

This is one explanation.

Hofberger, the Ford salesman in Tomcat, Va., a suburb of Washington, called up Littleton, of Littleton Menswear & Haberdashery, and said, "Good news, the new Fords have just come in and I've put one aside for you and your wife."

Littleton said, "I can't, Hofberger, my wife and I are getting a divorce."

"I'm sorry," Littleton said, "but I can't afford a new car this year. After I settle with my wife, I'll be lucky to buy a bicycle."

Hofberger hung up. His phone rang a few minutes later.

"This is Bedcheck the painter," the voice on the other end said. "When do you want us to start painting your house?"

"I changed my mind," said Hofberger, "I'm not going to paint the house."

"But I ordered the paint," Bedcheck said. "Why did you change your mind?"

"Because Littleton is getting a divorce and he can't afford a new car."

That evening when Bedcheck came home his wife said, "The new color television set arrived from Gladstone's TV Shop."

"Take it back," Bedcheck told his wife.

"Why?" she demanded.

"Because Hofberger isn't going to have his house painted now that the Littletons are getting a divorce."

The next day Mrs. Bedcheck dragged the TV set in its carton back to Gladstone. "We don't want it."

Gladstone's face dropped. He immediately called his travel agent, Sandstorm. "You know that trip you had scheduled for me to the Virgin Islands?"

"Right, the tickets are all written up."

"Cancel it. I can't go. Bedcheck just sent back the color TV set because Hofberger didn't sell a car to Littleton because they're going to get a divorce and she wants all his money."

Sandstorm tore up the airline tickets and went over to see his banker, Gripsholm. "I can't pay back the loan this month because Gladstone isn't going to the Virgin Islands."

Gripsholm was furious. When Rudemaker came in to borrow money for a new kitchen he needed for his restaurant, Gripsholm turned him down cold. "How can I loan you money when Sandstorm hasn't repaid the money he borrowed?"

Rudemaker called up the contractor, Eagleton, and said he couldn't put in a new kitchen. Eagleton laid off eight men.

Meanwhile, Ford announced it was giving a rebate on its new models. Hofberger called up Littleton immediately. "Good news," he said, "even if you are getting a divorce, you can afford a new car."

"I'm not getting a divorce," Littleton said. "It was all a misunderstanding and we've made up."

"That's great," Hofberger said. "Now you can buy the Ford."

"No way," said Littleton. "My business has been so lousy I don't know why I keep the doors open."

"I didn't realize that," Hofberger said.

"Do you realize I haven't seen Bedcheck, Gladstone, Sandstorm, Gripsholm, Rudemaker or Eagleton for more than a month? How can I stay in business if they don't patronize my store?"

Source: Art Buchwald, "Squaring the Economic Circle," *Cleveland Plain Dealer*, Feb. 22, 1975. Reprinted by permission.

SUMMARY

LO28.1 Describe how changes in income affect consumption (and saving).

Other things equal, a direct (positive) relationship exists between income and consumption and income and saving. The consumption and saving schedules show the various amounts that households intend to consume and save at the various income and output levels, assuming a fixed price level.

The *average* propensities to consume and save show the fractions of any total income that are consumed and saved; APC + APS = 1. The *marginal* propensities to consume and save show the fractions of any change in total income that are consumed and saved; MPC + MPS = 1.

LO28.2 List and explain factors other than income that can affect consumption.

The locations of the consumption and saving schedules (as they relate to real GDP) are determined by (*a*) the amount of wealth owned by households, (*b*) expectations of future prices and incomes, (*c*) real interest rates, (*d*) household debt, and (*e*) tax levels. The consumption and saving schedules are relatively stable.

LO28.3 Explain how changes in real interest rates affect investment.

The immediate determinants of investment are (*a*) the expected rate of return and (*b*) the real rate of interest. The economy's investment demand curve is found by cumulating investment projects, arraying them in descending order according to their expected rates of return, graphing the result, and applying the rule that investment should be undertaken up to the point at which the real interest rate, *i*, equals the expected rate of return, *r*. The investment demand curve reveals an inverse (negative) relationship between the interest rate and the level of aggregate investment.

Shifts of the investment demand curve can occur as the result of changes in (*a*) the acquisition, maintenance, and operating costs of capital goods; (*b*) business taxes; (*c*) technology; (*d*) the stocks of capital goods on hand; and (*e*) expectations.

Either changes in interest rates or shifts of the investment demand curve can change the level of investment.

LO28.4 Identify and explain factors other than the real interest rate that can affect investment.

The durability of capital goods, the irregular occurrence of major innovations, profit volatility, and the variability of expectations all contribute to the instability of investment spending.

LO28.5 Illustrate how changes in investment (or one of the other components of total spending) can increase or decrease real GDP by a multiple amount.

Through the multiplier effect, an increase in investment spending (or consumption spending, government purchases, or net export spending) ripples through the economy, ultimately creating a magnified increase in real GDP. The multiplier is the ultimate change in GDP divided by the initiating change in investment or some other component of spending.

The multiplier is equal to the reciprocal of the marginal propensity to save: The greater is the marginal propensity to save, the smaller is the multiplier. Also, the greater is the marginal propensity to consume, the larger is the multiplier.

Economists disagree on the size of the actual multiplier in the United States, with estimates ranging all the way from 2.5 to 0. But all estimates of actual-economy multipliers are less than the multiplier in our purposefully simple text illustrations.

TERMS AND CONCEPTS

45° (degree) line	average propensity to save (APS)	paradox of thrift
consumption schedule	marginal propensity to consume (MPC)	expected rate of return
saving schedule	marginal propensity to save (MPS)	investment demand curve
break-even income	wealth effect	multiplier
average propensity to consume (APC)		

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Precisely how do the MPC and the APC differ? How does the MPC differ from the MPS? Why must the sum of the MPC and the MPS equal 1? **LO28.1**
2. Why does a downshift of the consumption schedule typically involve an equal upshift of the saving schedule? What is the exception to this relationship? **LO28.2**
3. Why will a reduction in the real interest rate increase investment spending, other things equal? **LO28.3**
4. In what direction will each of the following occurrences shift the investment demand curve, other things equal? **LO28.4**
   a. An increase in unused production capacity occurs.
   b. Business taxes decline.
   c. The costs of acquiring equipment fall.
   d. Widespread pessimism arises about future business conditions and sales revenues.
   e. A major new technological breakthrough creates prospects for a wide range of profitable new products.
5. How is it possible for investment spending to increase even in a period in which the real interest rate rises? **LO28.4**
6. Why is investment spending unstable? **LO28.4**
7. Is the relationship between changes in spending and changes in real GDP in the multiplier effect a direct

(positive) relationship or is it an inverse (negative) relationship? How does the size of the multiplier relate to the size of the MPC? The MPS? What is the logic of the multiplier-MPC relationship? **LO28.5**

8. Why is the actual multiplier in the U.S. economy less than the multiplier in this chapter's example? **LO28.5**
9. **LAST WORD** What is the central economic idea humorously illustrated in Art Buchwald's piece "Squaring the Economic Circle"? How does the central idea relate to economic recessions, on the one hand, and vigorous economic expansions, on the other?

REVIEW QUESTIONS

1. What are the variables (the items measured on the axes) in a graph of the (*a*) consumption schedule and (*b*) saving schedule? Are the variables inversely (negatively) related or are they directly (positively) related? What is the fundamental reason that the levels of consumption and saving in the United States are each higher today than they were a decade ago? **LO28.1**
2. In year one, Adam earns $1,000 and saves $100. In year 2, Adam gets a $500 raise so that he earns a total of $1,500. Out of that $1,500, he saves $200. What is Adam's MPC out of his $500 raise? **LO28.1**
   - **a.** 0.50.
   - **b.** 0.75.
   - **c.** 0.80.
   - **d.** 1.00.
3. If the MPS rises, then the MPC will: **LO28.1**
   - **a.** Fall.
   - **b.** Rise.
   - **c.** Stay the same.
4. In what direction will each of the following occurrences shift the consumption and saving schedules, other things equal? **LO28.2**
   - **a.** A large decrease in real estate values, including private homes.
   - **b.** A sharp, sustained increase in stock prices.
   - **c.** A 5-year increase in the minimum age for collecting Social Security benefits.
   - **d.** An economy-wide expectation that a recession is over and that a robust expansion will occur.
   - **e.** A substantial increase in household borrowing to finance auto purchases.
5. Irving owns a chain of movie theaters. He is considering whether he should build a new theater downtown. The expected rate of return is 15 percent per year. He can borrow money at a 12 percent interest rate to finance the project. Should Irving proceed with this project? **LO28.3**
   - **a.** Yes.
   - **b.** No.
6. Which of the following scenarios will shift the investment demand curve right? **LO28.4**
   *Select **one or more** answers from the choices shown.*
   - **a.** Business taxes increase.
   - **b.** The expected return on capital increases.
   - **c.** Firms have a lot of unused production capacity.
   - **d.** Firms are planning on increasing their inventories.
7. True or False: Real GDP is more volatile (variable) than gross investment. **LO28.4**
8. If a $50 billion initial increase in spending leads to a $250 billion change in real GDP, how big is the multiplier? **LO28.5**
   - **a.** 1.0.
   - **b.** 2.5.
   - **c.** 4.0.
   - **d.** 5.0.
9. True or False: Larger MPCs imply larger multipliers. **LO28.5**

PROBLEMS

1. Refer to the table below. **LO28.1**
   - **a.** Fill in the missing numbers in the table.
   - **b.** What is the break-even level of income in the table? What is the term that economists use for the saving situation shown at the $240 level of income?
   - **c.** For each of the following items, indicate whether the value in the table is either constant or variable as income changes: the MPS, the APC, the MPC, the APS.
2. Suppose that disposable income, consumption, and saving in some country are $200 billion, $150 billion, and $50 billion,

Level of Output and Income (GDP = DI)	Consumption	Saving	APC	APS	MPC	MPS
$240	$______	$ −4	____	____		
					____	____
260	______	0	____	____		
					____	____
280	______	4	____	____		
					____	____
300	______	8	____	____		
					____	____
320	______	12	____	____		
					____	____
340	______	16	____	____		
					____	____
360	______	20	____	____		
					____	____
380	______	24	____	____		
					____	____
400	______	28	____	____		

respectively. Next, assume that disposable income increases by $20 billion, consumption rises by $18 billion, and saving goes up by $2 billion. What is the economy's MPC? Its MPS? What was the APC before the increase in disposable income? After the increase? **LO28.1**

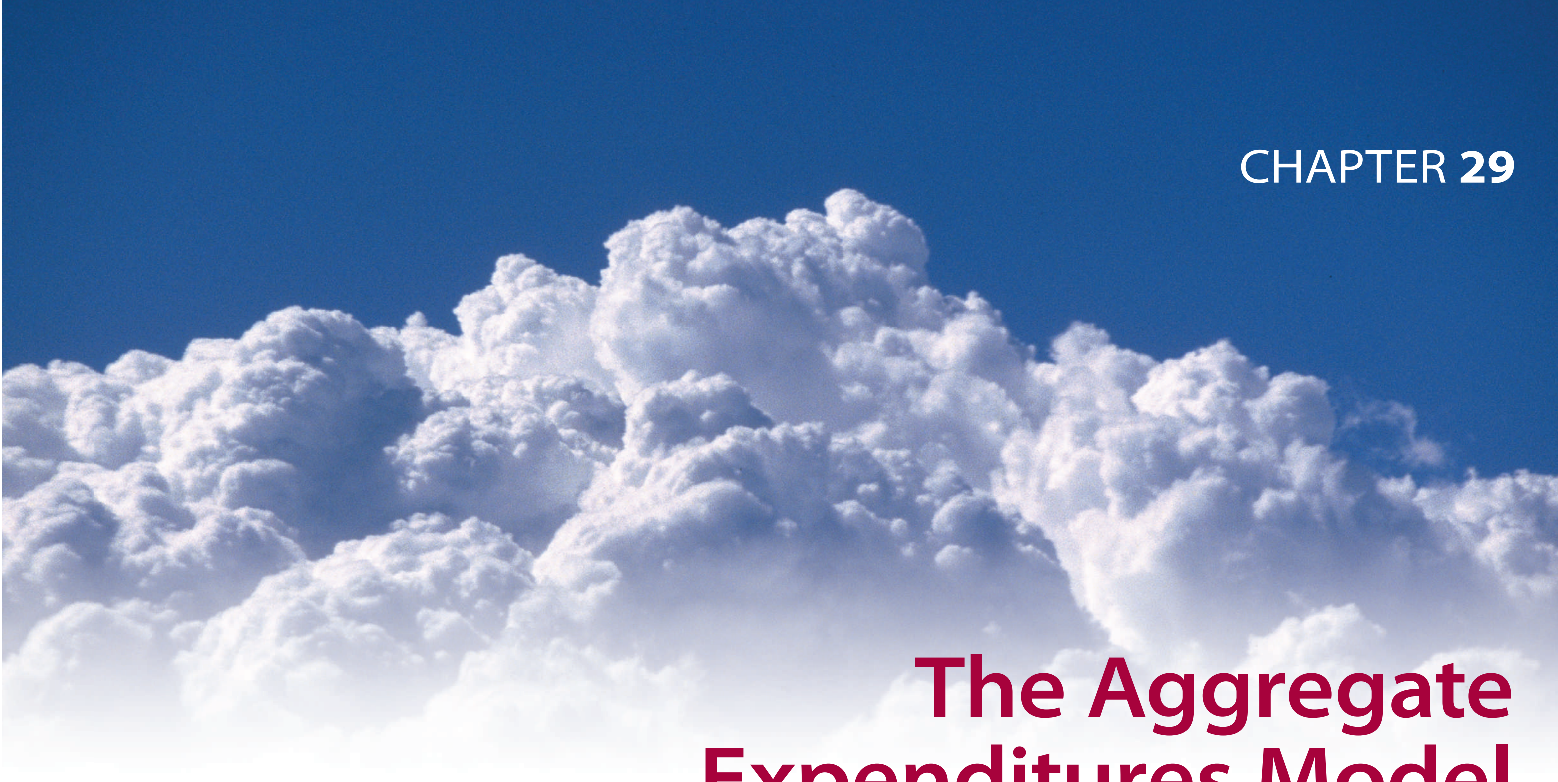
3. **ADVANCED ANALYSIS** Suppose that the linear equation for consumption in a hypothetical economy is $C = 40 + 0.8Y$. Also suppose that income (Y) is $400. Determine (*a*) the marginal propensity to consume, (*b*) the marginal propensity to save, (*c*) the level of consumption, (*d*) the average propensity to consume, (*e*) the level of saving, and (*f*) the average propensity to save. **LO28.1**
4. **ADVANCED ANALYSIS** Linear equations for the consumption and saving schedules take the general form $C = a + bY$ and $S = -a + (1 - b)Y$, where C, S, and Y are consumption, saving, and national income, respectively. The constant a represents the vertical intercept, and b represents the slope of the consumption schedule. **LO28.1, LO28.2**
   a. Use the following data to substitute numerical values for a and b in the consumption and saving equations.

National Income (Y)	Consumption (C)
$ 0	$ 80
100	140
200	200
300	260
400	320

   b. What is the economic meaning of b? Of $(1 - b)$?
   c. Suppose that the amount of saving that occurs at each level of national income falls by $20 but that the values of b and $(1 - b)$ remain unchanged. Restate the saving and consumption equations inserting the new numerical values, and cite a factor that might have caused the change.
5. Use your completed table for problem 1 to solve this problem. Suppose the wealth effect is such that $10 changes in wealth produce $1 changes in consumption at each level of income. If real estate prices tumble such that wealth declines by $80, what will be the new level of consumption and saving at the $340 billion level of disposable income? The new level of saving? **LO28.2**
6. Suppose a handbill publisher can buy a new duplicating machine for $500 and the duplicator has a 1-year life. The machine is expected to contribute $550 to the year's net revenue. What is the expected rate of return? If the real interest rate at which funds can be borrowed to purchase the machine is 8 percent, will the publisher choose to invest in the machine? Will it invest in the machine if the real interest rate is 9 percent? If it is 11 percent? **LO28.3**
7. Assume there are no investment projects in the economy that yield an expected rate of return of 25 percent or more. But suppose there are $10 billion of investment projects yielding expected returns of between 20 and 25 percent; another $10 billion yielding between 15 and 20 percent; another $10 billion between 10 and 15 percent; and so forth. Cumulate these data and present them graphically, putting the expected rate of return (and the real interest rate) on the vertical axis and the amount of investment on the horizontal axis. What will be the equilibrium level of aggregate investment if the real interest rate is (*a*) 15 percent, (*b*) 10 percent, and (*c*) 5 percent? **LO28.3**
8. Refer to the table in Figure 28.5 and suppose that the real interest rate is 6 percent. Next, assume that some factor changes such that the expected rate of return declines by 2 percentage points at each prospective level of investment. Assuming no change in the real interest rate, by how much and in what direction will investment change? Which of the following might cause this change: (*a*) a decision to increase inventories; (*b*) an increase in excess production capacity? **LO28.4**
9. What will the multiplier be when the MPS is 0, 0.4, 0.6, and 1? What will it be when the MPC is 1, 0.90, 0.67, 0.50, and 0? How much of a change in GDP will result if firms increase their level of investment by $8 billion and the MPC is 0.80? If the MPC instead is 0.67? **LO28.5**
10. Suppose that an initial $10 billion increase in investment spending expands GDP by $10 billion in the first round of the multiplier process. If GDP and consumption both rise by $6 billion in the second round of the process, what is the MPC in this economy? What is the size of the multiplier? If, instead, GDP and consumption both rose by $8 billion in the second round, what would have been the size of the multiplier? **LO28.5**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.



CHAPTER **29**

The Aggregate Expenditures Model

Learning Objectives

LO29.1 Explain how sticky prices relate to the aggregate expenditures model.

LO29.2 Explain how an economy's investment schedule is derived from the investment demand curve and an interest rate.

LO29.3 Illustrate how economists combine consumption and investment to depict an aggregate expenditures schedule for a private closed economy and how that schedule can be used to demonstrate the economy's equilibrium level of output (where the total quantity of goods produced equals the total quantity of goods purchased).

LO29.4 Discuss the two other ways to characterize the equilibrium level of real GDP in a private closed economy: saving = investment, and no unplanned changes in inventories.

LO29.5 Analyze how changes in equilibrium real GDP can occur in the aggregate expenditures model and describe how those changes relate to the multiplier.

LO29.6 Explain how economists integrate the international sector (exports and imports) into the aggregate expenditures model.

LO29.7 Explain how economists integrate the public sector (government expenditures and taxes) into the aggregate expenditures model.

LO29.8 Differentiate between equilibrium GDP and full-employment GDP and identify and describe the nature and causes of "recessionary expenditure gaps" and "inflationary expenditure gaps."

In previous chapters we answered in detail two of the most critical questions in macroeconomics: How is an economy's output measured? Why does an economy grow? But we have been relatively general in addressing two other important questions: What determines the level of GDP, given a nation's production capacity? What causes real GDP to rise in one period and to fall in another? To provide more thorough answers to these two questions, we construct the aggregate expenditures model, which has its origins in 1936 in the writings of British economist John Maynard Keynes (pronounced "Caines"). The basic premise of the aggregate expenditures model—also known as the "Keynesian cross" model—is that the amount of goods and services produced and therefore the level of employment depend directly on the level of aggregate expenditures (total spending). Businesses will produce only a level of output that they think they can profitably sell. They will idle their workers and machinery when there are no markets for their goods and services.

ORIGIN OF THE IDEA

O29.1
Aggregate expenditures model

Assumptions and Simplifications

LO29.1 Explain how sticky prices relate to the aggregate expenditures model.

The simplifying assumptions underpinning the aggregate expenditures model reflect the economic conditions that were prevalent during the Great Depression. As discussed in this chapter's Last Word, Keynes created the model during the middle of the Great Depression in the hopes of understanding both why the Great Depression had happened as well as how it might be ended.

The most fundamental assumption behind the aggregate expenditures model is that prices in the economy are fixed. In the terminology of Chapter 24, the aggregate expenditures model is an extreme version of a sticky price model. In fact, it is a stuck-price model because the price-level cannot change at all.

Keynes made this simplifying assumption because he had observed that prices had not declined sufficiently during the Great Depression to boost spending and maintain output and employment at their pre-Depression levels. Such price declines had been predicted by macroeconomic theories that were popular before the Great Depression. But actual prices did not fall sufficiently during the Great Depression, and the economy sank far below its potential output. Real GDP in the United States declined by 27 percent from 1929 to 1933, and the unemployment rate rose to 25 percent. Thousands of factories sat idle, gathering dust and producing nothing, because nobody wanted to buy their output.

To Keynes, this massive unemployment of labor and capital resulted from firms reacting to information about how much they should produce. As households and businesses greatly reduced their spending, inventories of unsold goods rocketed. Unable or unwilling to slash their prices, firms could not sell all the goods they had already produced. So they greatly reduced their current production. This meant discharging workers, idling production lines, and even closing entire factories. Keynes thought that a new economic model was needed to show how all this could have happened and how it might be reversed.

The Keynesian aggregate expenditures model is not just of historical interest. It is still insightful even today because many prices in the modern economy are inflexible downward over relatively short periods of time. The aggregate expenditures model therefore can help us understand how the modern economy is likely to initially adjust to various economic shocks over shorter periods of time. For example, it clarifies aspects of the severe 2007–2009 recession, such as why unexpected initial declines in spending caused even larger declines in real GDP. It also illuminates the thinking underlying the stimulus programs (tax cuts, government spending increases) enacted by the government during the recession.

We will build up the aggregate expenditures model in simple stages. Let's first look at aggregate expenditures and equilibrium GDP in a *private closed economy*—one without international trade or government. Then we will "open" the "closed" economy to exports and imports and also convert our "private" economy to a more realistic "mixed" economy that includes government purchases (or, more loosely, "government spending") and taxes.

In addition, until we introduce taxes into the model, we will assume that real GDP equals disposable income (DI). For instance, if $500 billion of output is produced as GDP, households will receive exactly $500 billion of disposable income that they can then consume or save. And finally, unless specified otherwise, we will assume (as Keynes did) that the presence of excess production capacity and unemployed labor implies that an increase in aggregate expenditures will increase real output and employment without raising the price level.

Consumption and Investment Schedules

LO29.2 Explain how an economy's investment schedule is derived from the investment demand curve and an interest rate.

In the private closed economy, the two components of aggregate expenditures are consumption, C, and gross investment, I_g. Because we examined the *consumption schedule* (Figure 28.2a) in the previous chapter, there is no need to repeat that analysis here. But to add the investment decisions of businesses to the consumption plans of households, we need to construct an investment schedule showing the amounts business firms collectively intend to invest—their **planned investment**—at each possible level of GDP. Such a schedule represents the investment plans of businesses in the same way the consumption schedule represents the consumption plans of households. In developing the investment schedule, we will assume that this planned investment is independent of the level of current disposable income or real output.

Suppose the investment demand curve is as shown in Figure 29.1a and the current real interest rate is 8 percent. This means that firms will spend $20 billion on investment goods. Our assumption tells us that this $20 billion of investment will occur at both low and high levels of GDP. The line I_g in Figure 29.1b shows this graphically;

FIGURE 29.1 (a) The investment demand curve and (b) the investment schedule. (a) The level of investment spending (here, $20 billion) is determined by the real interest rate (here, 8 percent) together with the investment demand curve *ID*. (b) The investment schedule I_g relates the amount of investment ($20 billion) determined in (a) to the various levels of GDP.

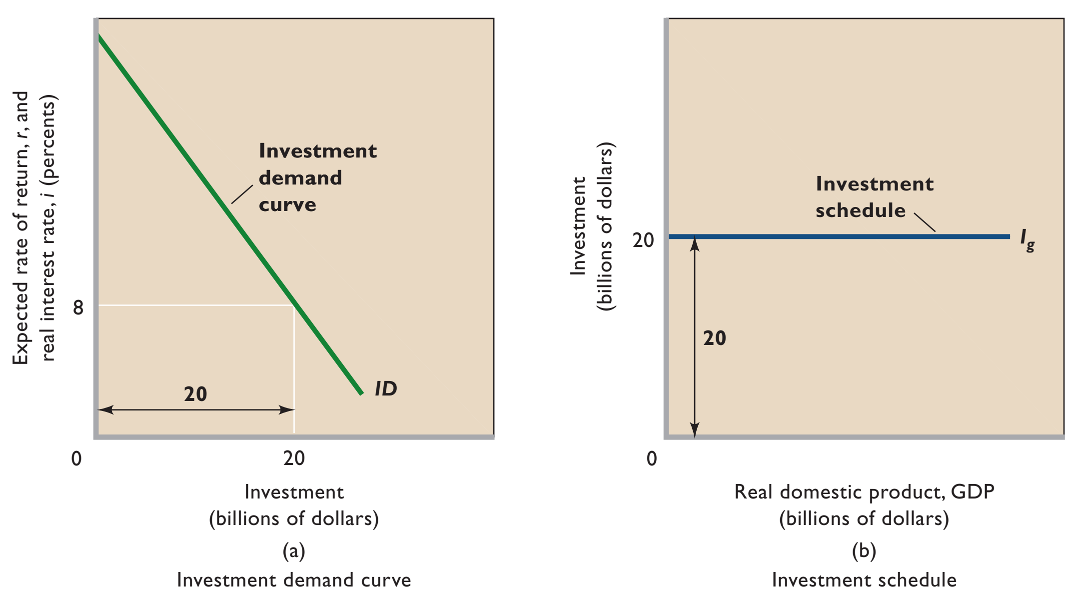


TABLE 29.1 **The Investment Schedule (in Billions)**

(1) Level of Real Output and Income	(2) Investment (I_g)
$370	$20
390	20
410	20
430	20
450	20
470	20
490	20
510	20
530	20
550	20

it is the economy's **investment schedule.** You should not confuse this investment schedule I_g with the investment demand curve *ID* in Figure 29.1a. The investment schedule shows the amount of investment forthcoming at each level of GDP. As indicated in Figure 29.1b, the interest rate and investment demand curve together determine this amount ($20 billion). Table 29.1 shows the investment schedule in tabular form. Note that investment (I_g) in column 2 is $20 billion at all levels of real GDP.

Equilibrium GDP: $C + I_g$ = GDP

LO29.3 Illustrate how economists combine consumption and investment to depict an aggregate expenditures schedule for a private closed economy and how that schedule can be used to demonstrate the economy's equilibrium level of output (where the total quantity of goods produced equals the total quantity of goods purchased).

Now let's combine the consumption schedule of Chapter 28 and the investment schedule here to explain the equilibrium levels of output, income, and employment in the private closed economy.

Tabular Analysis

Columns 2 through 5 in Table 29.2 repeat the consumption and saving schedules of Table 28.1 and the investment schedule of Table 29.1.

Real Domestic Output Column 2 in Table 29.2 lists the various possible levels of total output—of real GDP—that the private sector might produce. Firms would be willing to produce any one of these 10 levels of output just as long as the revenue that they receive from selling any particular level equals or exceeds the costs they would incur to produce it. Those costs are the factor payments needed to obtain the required amounts of land, labor, capital, and entrepreneurship. For example, firms would be willing to produce $370 billion of output if the costs of production (wages, rents, interest, and the normal profit needed to attract entrepreneurship) are less than or equal to the $370 billion in revenue that they would get from selling the output.

Aggregate Expenditures In the private closed economy of Table 29.2, aggregate expenditures consist of consumption (column 3) plus investment (column 5). Their

TABLE 29.2 **Determination of the Equilibrium Levels of Employment, Output, and Income: A Private Closed Economy**

(1) Possible Levels of Employment, Millions	(2) Real Domestic Output (and Income) (GDP = DI),* Billions	(3) Consumption (C), Billions	(4) Saving (S), Billions	(5) Investment (I_g), Billions	(6) Aggregate Expenditures ($C + I_g$), Billions	(7) Unplanned Changes in Inventories, (+ or −)	(8) Tendency of Employment, Output, and Income
(1) 40	$370	$375	$−5	$20	$395	$−25	Increase
(2) 45	390	390	0	20	410	−20	Increase
(3) 50	410	405	5	20	425	−15	Increase
(4) 55	430	420	10	20	440	−10	Increase
(5) 60	450	435	15	20	455	−5	Increase
(6) **65**	**470**	**450**	**20**	**20**	**470**	**0**	**Equilibrium**
(7) 70	490	465	25	20	485	+5	Decrease
(8) 75	510	480	30	20	500	+10	Decrease
(9) 80	530	495	35	20	515	+15	Decrease
(10) 85	550	510	40	20	530	+20	Decrease

*If depreciation and net foreign factor income are zero, government is ignored, and it is assumed that all saving occurs in the household sector of the economy, then GDP as a measure of domestic output is equal to NI, PI, and DI. This means that households receive a DI equal to the value of total output.

sum is shown in column 6, which along with column 2 makes up the **aggregate expenditures schedule** for the private closed economy. This schedule shows the amount ($C + I_g$) that will be spent at each possible output or income level.

At this point we are working with *planned investment*—the data in column 5, Table 29.2. These data show the amounts firms plan or intend to invest, not the amounts they actually will invest if there are unplanned changes in inventories. More about that shortly.

Equilibrium GDP Of the 10 possible levels of GDP in Table 29.2, which is the equilibrium level? Which total output is the economy capable of sustaining?

The equilibrium output is that output whose production creates total spending just sufficient to purchase that output. So the equilibrium level of GDP is the level at which the total quantity of goods produced (GDP) equals the total quantity of goods purchased ($C + I_g$). In the private closed economy, the **equilibrium GDP** is where

$$C + I_g = \text{GDP}$$

If you look at the domestic output levels in column 2 and the aggregate expenditures levels in column 6, you will see that this equality exists only at $470 billion of GDP (row 6). That is the only output at which economy-wide spending is precisely equal to the amount needed to move that output off the shelves. At $470 billion of GDP, the annual rates of production and spending are in balance. There is no overproduction, which would result in a piling up of unsold goods and consequently cutbacks in the production rate. Nor is there an excess of total spending, which would draw down inventories of goods and prompt increases in the rate of production. In short, there is no reason for businesses to alter this rate of production; $470 billion is the equilibrium GDP.

Disequilibrium No level of GDP other than the equilibrium level of GDP can be sustained. At levels of GDP *less than* equilibrium, spending always exceeds GDP. If, for example, firms produced $410 billion of GDP (row 3 in Table 29.2), they would find it would yield $405 billion in consumer spending. Supplemented by $20 billion of planned investment, aggregate expenditures ($C + I_g$) would be $425 billion, as shown in column 6. The economy would provide an annual rate of spending more than sufficient to purchase the $410 billion of annual production. Because buyers would be taking goods off the shelves faster than firms could produce them, an unplanned decline in business inventories of $15 billion would occur (column 7) if this situation continued. But businesses can adjust to such an imbalance between aggregate expenditures and real output by stepping up production. Greater output will increase employment and total income. This process will continue until the equilibrium level of GDP is reached ($470 billion).

The reverse is true at all levels of GDP *greater than* the $470 billion equilibrium level. Businesses will find that these total outputs fail to generate the spending needed to clear the shelves of goods. Being unable to recover their costs, businesses will cut back on production. To illustrate: At the $510 billion output (row 8), business managers would find spending is insufficient to permit the sale of all that output. Of the $510 billion of income that this output creates, $480 billion would be received back by businesses as consumption spending. Though supplemented by $20 billion of planned investment spending, total expenditures ($500 billion) would still be $10 billion below the $510 billion quantity produced. If this imbalance persisted, $10 billion of inventories would pile up (column 7). But businesses can adjust to this unintended accumulation of unsold goods by cutting back on the rate of production. The resulting decline in output would mean fewer jobs and a decline in total income.

WORKED PROBLEMS

W29.1
Equilibrium GDP

Graphical Analysis

We can demonstrate the same analysis graphically. In **Figure 29.2 (Key Graph)** the 45° line developed in Chapter 28 now takes on increased significance. Recall that at any point on this line, the value of what is being measured on the horizontal axis (here, GDP) is equal to the value of what is being measured on the vertical axis (here, aggregate expenditures, or $C + I_g$). Having discovered in our tabular analysis that the equilibrium level of domestic output is determined where $C + I_g$ equals GDP, we can say that the 45° line in Figure 29.2 is a graphical statement of that equilibrium condition.

Now we must graph the aggregate expenditures schedule onto Figure 29.2. To do this, we duplicate the consumption schedule C in Figure 28.2a and add to it vertically the constant $20 billion amount of investment I_g from Figure 29.1b. This $20 billion is the amount we assumed firms plan to invest at all levels of GDP. Or, more directly, we can plot the $C + I_g$ data in column 6, Table 29.2.

Observe in Figure 29.2 that the aggregate expenditures line $C + I_g$ shows that total spending rises with income and

KEY GRAPH

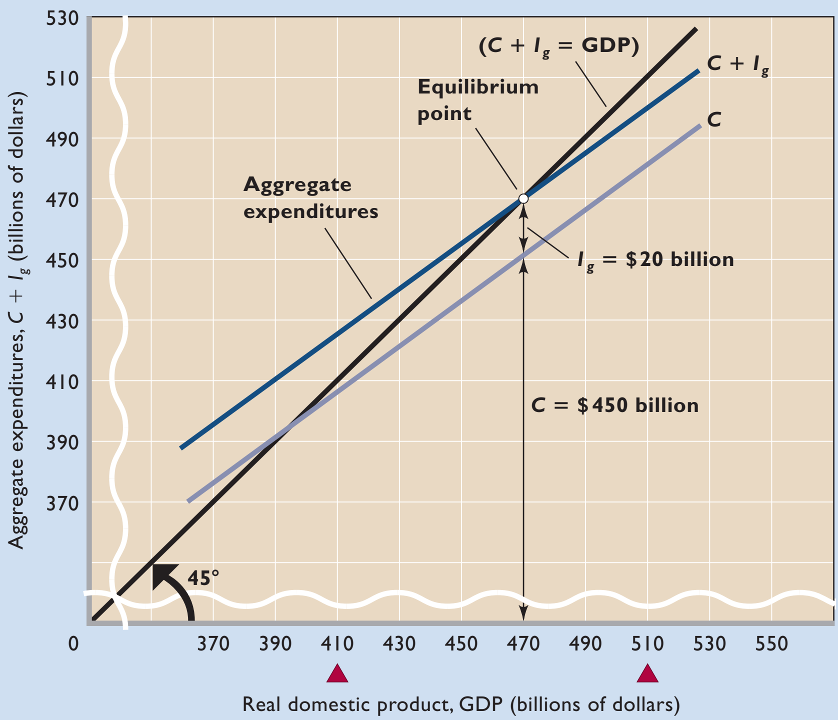


FIGURE 29.2 Equilibrium GDP in a private closed economy. The aggregate expenditures schedule, $C + I_g$, is determined by adding the investment schedule I_g to the upsloping consumption schedule C. Since investment is assumed to be the same at each level of GDP, the vertical distances between C and $C + I_g$ do not change. Equilibrium GDP is determined where the aggregate expenditures schedule intersects the 45° line, in this case at $470 billion.

QUICK QUIZ FOR FIGURE 29.2

1. In this figure, the slope of the aggregate expenditures schedule $C + I_g$:
  a. increases as real GDP increases.
  b. falls as real GDP increases.
  c. is constant and equals the MPC.
  d. is constant and equals the MPS.

2. At all points on the 45° line:
  a. equilibrium GDP is possible.
  b. aggregate expenditures exceed real GDP.
  c. consumption exceeds investment.
  d. aggregate expenditures are less than real GDP.

3. The $490 billion level of real GDP is not at equilibrium because:
  a. investment exceeds consumption.
  b. consumption exceeds investment.
  c. planned $C + I_g$ exceeds real GDP.
  d. planned $C + I_g$ is less than real GDP.

4. The $430 billion level of real GDP is not at equilibrium because:
  a. investment exceeds consumption.
  b. consumption exceeds investment.
  c. planned $C + I_g$ exceeds real GDP.
  d. planned $C + I_g$ is less than real GDP.

Answers: 1. c; 2. a; 3. d; 4. c

output (GDP), but not as much as income rises. That is true because the marginal propensity to consume—the slope of line C—is less than 1. A part of any increase in income will be saved rather than spent. And because the aggregate expenditures line $C + I_g$ is parallel to the consumption line C, the slope of the aggregate expenditures line also equals the MPC for the economy and is less than 1. For our particular data, aggregate expenditures rise by $15 billion for every $20 billion increase in real output and income because $5 billion of each $20 billion increment is saved. Therefore, the slope of the aggregate expenditures line is 0.75 (= Δ$15/Δ$20).

The equilibrium level of GDP is determined by the intersection of the aggregate expenditures schedule and the 45° line. This intersection locates the only point at which aggregate expenditures (on the vertical axis) are equal to GDP (on the horizontal axis). Because Figure 29.2 is based on the data in Table 29.2, we once again find that equilibrium output is $470 billion. Observe that consumption at this output is $450 billion and investment is $20 billion.

It is evident from Figure 29.2 that no levels of GDP *above* the equilibrium level are sustainable because at those levels $C + I_g$ falls short of GDP. Graphically, the aggregate expenditures schedule lies below the 45° line in those situations. At the \$510 billion GDP level, for example, $C + I_g$ is only \$500 billion. This underspending causes inventories to rise, prompting firms to readjust production downward, in the direction of the \$470 billion output level.

Conversely, at levels of GDP *below* \$470 billion, the economy wants to spend in excess of what businesses are producing. Then $C + I_g$ exceeds total output. Graphically, the aggregate expenditures schedule lies above the 45° line. At the \$410 billion GDP level, for example, $C + I_g$ totals \$425 billion. This excess spending causes inventories to fall below their planned level, prompting firms to adjust production upward, in the direction of the \$470 billion output level. Once production reaches that level, it will be sustained there indefinitely unless there is some change in the location of the aggregate expenditures line.

Other Features of Equilibrium GDP

LO29.4 Discuss the two other ways to characterize the equilibrium level of real GDP in a private closed economy: saving = investment, and no unplanned changes in inventories.

We have seen that $C + I_g =$ GDP at equilibrium in the private closed economy. A closer look at Table 29.2 reveals two more characteristics of equilibrium GDP:

- Saving and *planned* investment are equal ($S = I_g$).
- There are no *unplanned* changes in inventories.

Saving Equals Planned Investment

As shown by row 6 in Table 29.2, saving and planned investment are both \$20 billion at the \$470 billion equilibrium level of GDP.

Saving is a **leakage** or withdrawal of spending from the economy's circular flow of income and expenditures. Saving is what causes consumption to be less than total output or GDP. Because of saving, consumption by itself is insufficient to remove domestic output from the shelves, apparently setting the stage for a decline in total output.

However, firms do not intend to sell their entire output to consumers. Some of that output will be capital goods sold to other businesses. Investment—the purchases of capital goods—is therefore an **injection** of spending into the income-expenditures stream. As an adjunct to consumption, investment is thus a potential replacement for the leakage of saving.

If the leakage of saving at a certain level of GDP exceeds the injection of investment, then $C + I_g$ will be less than GDP and that level of GDP cannot be sustained. Any GDP for which saving exceeds investment is an above-equilibrium GDP. Consider GDP of \$510 billion (row 8 in Table 29.2). Households will save \$30 billion, but firms will plan to invest only \$20 billion. This \$10 billion excess of saving over planned investment will reduce total spending to \$10 billion below the value of total output. Specifically, aggregate expenditures will be \$500 billion while real GDP is \$510 billion. This spending deficiency will reduce real GDP.

Conversely, if the injection of investment exceeds the leakage of saving, then $C + I_g$ will be greater than GDP and drive GDP upward. Any GDP for which investment exceeds saving is a below-equilibrium GDP. For example, at a GDP of \$410 billion (row 3 in Table 29.2), households will save only \$5 billion, but firms will invest \$20 billion. So investment exceeds saving by \$15 billion. The small leakage of saving at this relatively low GDP level is more than compensated for by the larger injection of investment spending. That causes $C + I_g$ to exceed GDP and drives GDP higher.

Only where $S = I_g$—where the leakage of saving of \$20 billion is exactly offset by the injection of planned investment of \$20 billion—will aggregate expenditures ($C + I_g$) equal real output (GDP). That $C + I_g =$ GDP equality is what defines the equilibrium GDP.

No Unplanned Changes in Inventories

As part of their investment plans, firms may decide to increase or decrease their inventories. But, as confirmed in line 6 of Table 29.2, there are no **unplanned changes in inventories** at equilibrium GDP. This fact, along with $C + I_g =$ GDP, and $S = I_g$ is a characteristic of equilibrium GDP in the private closed economy.

Unplanned changes in inventories play a major role in achieving equilibrium GDP. Consider, as an example, the \$490 billion *above-equilibrium* GDP shown in row 7 of Table 29.2. What happens if firms produce that output, thinking they can sell it? Households save \$25 billion of their \$490 billion DI, so consumption is only \$465 billion. Planned investment—which includes *planned* changes in inventories—is \$20 billion (column 5). So aggregate expenditures ($C + I_g$) are \$485 billion and sales fall short of production by \$5 billion. Firms retain that extra \$5 billion of goods as an *unplanned* increase in inventories (column 7). It results from the failure of total spending to remove total output from the shelves.

Because changes in inventories are a part of investment, we note that *actual investment* is \$25 billion. It consists of \$20 billion of planned investment *plus* the \$5 billion unplanned increase in inventories. Actual investment equals the saving of \$25 billion, even though saving exceeds planned investment by $5 billion. Because firms cannot earn profits by accumulating unwanted inventories, the $5 billion unplanned increase in inventories will prompt them to cut back employment and production. GDP will fall to its equilibrium level of $470 billion, at which unplanned changes in inventories are zero.

Now look at the *below-equilibrium* $450 billion output (row 5, Table 29.2). Because households save only $15 billion of their $450 billion DI, consumption is $435 billion. Planned investment by firms is $20 billion, so aggregate expenditures are $455 billion. Sales exceed production by $5 billion. This is so only because a $5 billion unplanned decrease in business inventories has occurred. Firms must *disinvest* $5 billion in inventories (column 7). Note again that actual investment is $15 billion ($20 billion planned *minus* the $5 billion decline in inventory investment) and is equal to saving of $15 billion, even though planned investment exceeds saving by $5 billion. The unplanned decline in inventories, resulting from the excess of sales over production, will encourage firms to expand production. GDP will rise to $470 billion, at which unplanned changes in inventories are zero.

When economists say differences between investment and saving can occur and bring about changes in equilibrium GDP, they are referring to planned investment and saving. Equilibrium occurs only when planned investment and saving are equal. But when unplanned changes in inventories are considered, investment and saving are always equal, regardless of the level of GDP. That is true because actual investment consists of planned investment and unplanned investment (unplanned changes in inventories). Unplanned changes in inventories act as a balancing item that equates the actual amounts saved and invested in any period.

Changes in Equilibrium GDP and the Multiplier

LO29.5 Analyze how changes in equilibrium real GDP can occur in the aggregate expenditures model and describe how those changes relate to the multiplier.

In the previous chapter, we established that an initial change in spending can cause a greater change in real output through the multiplier effect. In equation form,

$$\text{Multiplier} = \frac{\text{change in real GDP}}{\text{initial change in spending}}$$

Further, we discovered that the size of the multiplier depends on the size of the MPS in the economy:

$$\text{Multiplier} = \frac{1}{\text{MPS}}$$

(Because the multiplier is such an important element of the aggregate expenditures model, we highly recommend that you quickly review Figure 28.8 at this time.)

In a private closed economy, the equilibrium GDP will change in response to changes in either the investment schedule or the consumption schedule. Because changes in the investment schedule are the main sources of instability, we will direct our attention toward them.

Figure 29.3 shows the effect of changes in investment spending on the equilibrium real GDP. Suppose that the expected rate of return on investment rises or that the real interest rate falls such that investment spending increases by $5 billion. That would be shown as an upward shift of the investment schedule in Figure 29.1b. In Figure 29.3, the $5 billion increase of investment spending will increase aggregate expenditures from $(C + I_g)_0$ to $(C + I_g)_1$ and raise equilibrium real GDP from $470 billion to $490 billion.

If the expected rate of return on investment decreases or if the real interest rate rises, investment spending will decline by, say, $5 billion. That would be shown as a downward shift of the investment schedule in Figure 29.1b and a downward shift of the aggregate expenditures schedule from $(C + I_g)_0$ to $(C + I_g)_2$ in Figure 29.3. Equilibrium GDP will fall from $470 billion to $450 billion.

In our examples, a $5 billion change in investment spending leads to a $20 billion change in output and income. So the *multiplier* is 4 (= $20/$5). The MPS is 0.25, meaning that for every $1 billion of new income, $0.25 billion of new saving occurs. Therefore, $20 billion of new income is needed to generate $5 billion of new saving. Once that increase in income and saving occurs, the economy is back in equilibrium—$C + I_g$ = GDP; saving and investment are equal; and there are no unplanned changes in inventories. You can see, then, why the multiplier is equal to 1/MPS and that the multiplier process is an integral part of the aggregate expenditures model.

QUICK REVIEW 29.1

- In a private closed economy, equilibrium GDP occurs where aggregate expenditures equal real domestic output ($C + I_g$ = GDP).
- At equilibrium GDP, saving equals planned investment ($S = I_g$) and unplanned changes in inventories are zero.
- Actual investment consists of planned investment plus unplanned changes in inventories (+ or −) and is always equal to saving in a private closed economy.
- Through the multiplier effect, an initial change in investment spending can cause a magnified change in domestic output and income.

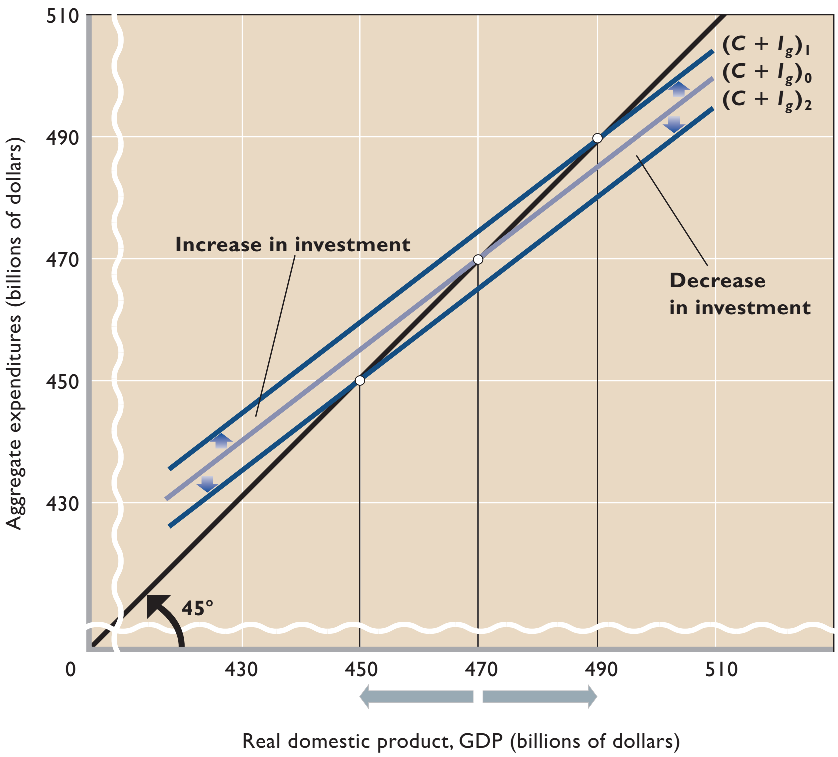


FIGURE 29.3 Changes in the aggregate expenditures schedule and the multiplier effect. An upward shift of the aggregate expenditures schedule from $(C + I_g)_0$ to $(C + I_g)_1$ will increase the equilibrium GDP. Conversely, a downward shift from $(C + I_g)_0$ to $(C + I_g)_2$ will lower the equilibrium GDP. The extent of the changes in equilibrium GDP will depend on the size of the multiplier, which in this case is 4 (= 20/5). The multiplier is equal to 1/MPS (here, 4 = 1/.25).

Adding International Trade

LO29.6 Explain how economists integrate the international sector (exports and imports) into the aggregate expenditures model.

We next move from a private closed economy to a private open economy that incorporates exports (X) and imports (M). Our focus will be on **net exports** (exports minus imports), which may be either positive or negative.

Net Exports and Aggregate Expenditures

Like consumption and investment, exports create domestic production, income, and employment for a nation. Although U.S. goods and services produced for export are sent abroad, foreign spending on those goods and services increases production and creates jobs and incomes in the United States. We must therefore include exports as a component of U.S. aggregate expenditures.

Conversely, when an economy is open to international trade, it will spend part of its income on imports—goods and services produced abroad. To avoid overstating the value of domestic production, we must subtract the amount spent on imported goods because such spending generates production and income abroad rather than at home. So, to correctly measure aggregate expenditures for domestic goods and services, we must subtract expenditures on imports from total spending.

In short, for a private closed economy, aggregate expenditures are $C + I_g$. But for an open economy, aggregate expenditures are $C + I_g + (X - M)$. Or, recalling that net exports (X_n) equal $(X - M)$, we can say that aggregate expenditures for a private open economy are $C + I_g + X_n$.

The Net Export Schedule

A net export schedule lists the amount of net exports that will occur at each level of GDP. Table 29.3 shows two possible net export schedules for the hypothetical economy represented in Table 29.2. In net export schedule X_{n1}

TABLE 29.3 Two Net Export Schedules (in Billions)

(1) Level of GDP	(2) Net Exports, X_{n1} ($X > M$)	(3) Net Exports, X_{n2} ($X < M$)
$370	$+5	$−5
390	+5	−5
410	+5	−5
430	+5	−5
450	+5	−5
470	+5	−5
490	+5	−5
510	+5	−5
530	+5	−5
550	+5	−5

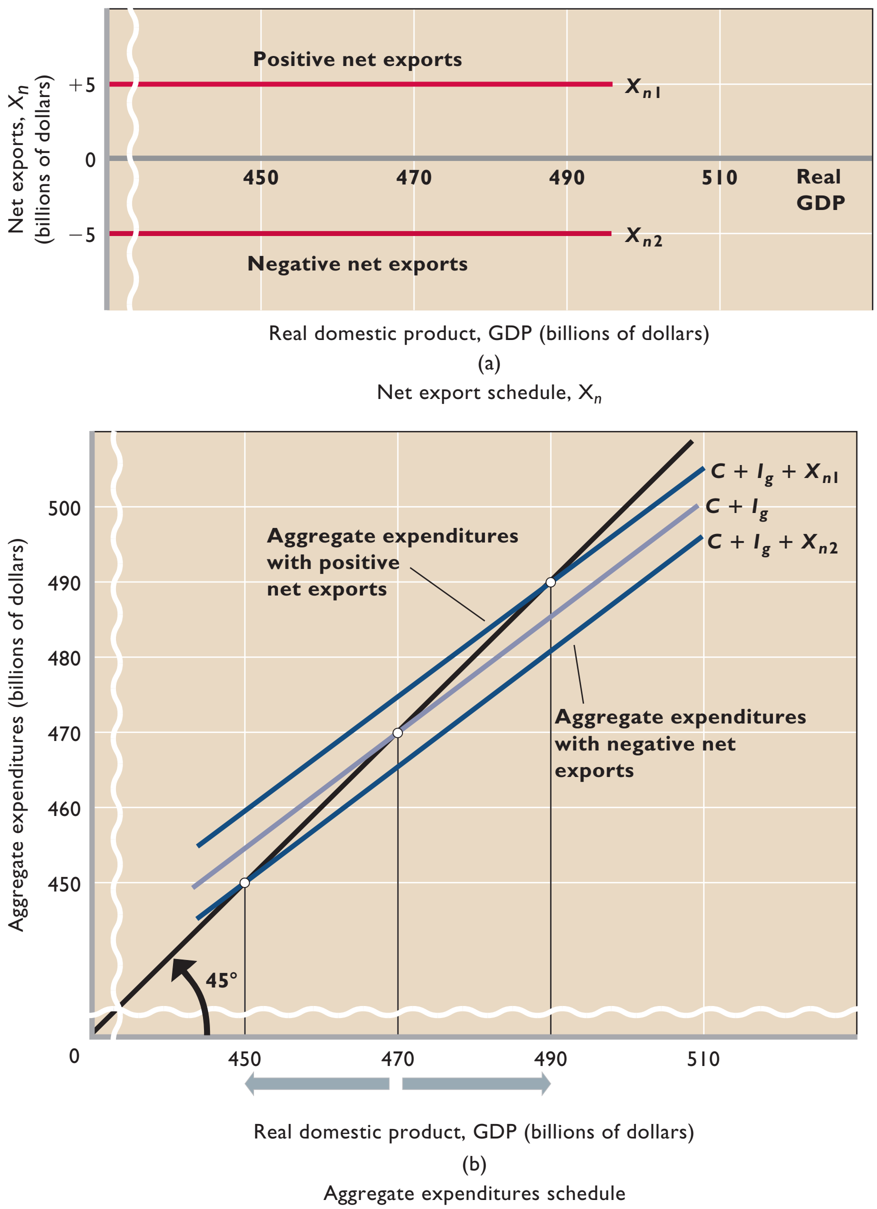


FIGURE 29.4 Net exports and equilibrium GDP. (a) Net exports can be either positive, as shown by the net export schedule X_{n1} or negative, as depicted by net export schedule X_{n2}. (b) Positive net exports elevate the aggregate expenditure schedule from the closed-economy level of $C + I_g$ to the open-economy level of $C + I_g + X_{n1}$. Negative net exports lower the aggregate expenditures schedule from the closed-economy level of $C + I_g$ to the open-economy level of $C + I_g + X_{n2}$.

(columns 1 and 2), exports exceed imports by \$5 billion at each level of GDP. Perhaps exports are \$15 billion while imports are \$10 billion. In schedule X_{n2} (columns 1 and 3), imports are \$5 billion higher than exports. Perhaps imports are \$20 billion while exports are \$15 billion. To simplify our discussion, we assume in both schedules that net exports are independent of GDP.[1]

Figure 29.4a represents the two net export schedules in Table 29.3. Schedule X_{n1} is above the horizontal axis and depicts positive net exports of \$5 billion at all levels of GDP. Schedule X_{n2}, which is below the horizontal axis, shows negative net exports of \$5 billion at all levels of GDP.

Net Exports and Equilibrium GDP

The aggregate expenditures schedule labeled $C + I_g$ in Figure 29.4b reflects the private closed economy. It shows the combined consumption and gross investment expenditures occurring at each level of GDP. With no foreign sector, the equilibrium GDP is \$470 billion.

[1]In reality, although our exports depend on foreign incomes and are thus independent of U.S. GDP, our imports do vary directly with our own domestic national income. Just as our domestic consumption varies directly with our GDP, so do our purchases of foreign goods. As our GDP rises, U.S. households buy not only more Cadillacs and Harleys, but also more Mercedes and Kawasakis. However, for now we will ignore the complications of the positive relationship between imports and U.S. GDP.

But in the private open economy, net exports can be either positive or negative. Let's see how each of the net export schedules in Figure 29.4a affects equilibrium GDP.

Positive Net Exports Suppose the net export schedule is X_{n1}. The \$5 billion of additional net export expenditures by the rest of the world is accounted for by adding that \$5 billion to the $C + I_g$ schedule in Figure 29.4b. Aggregate expenditures at each level of GDP are then \$5 billion higher than $C + I_g$ alone. The aggregate expenditures schedule for the open economy thus becomes $C + I_g + X_{n1}$. In this case, international trade increases equilibrium GDP from \$470 billion in the private closed economy to \$490 billion in the private open economy. Adding net exports of \$5 billion has increased GDP by \$20 billion, in this case implying a multiplier of 4.

Generalization: Other things equal, positive net exports increase aggregate expenditures and GDP beyond what they would be in a closed economy. Be careful to notice that this increase is the result of exports being larger than imports. This is true because exports and imports have opposite effects on the measurement of domestically produced output. Exports increase real GDP by increasing expenditures on domestically produced output. Imports, by contrast, must be subtracted when calculating real GDP because they are expenditures directed toward output produced abroad. It is only because net exports are positive in this example—so that the expansionary effect of exports outweighs the reductions caused by imports—that we get the overall increase in real GDP. As the next section shows, if net exports are negative, then the reductions caused by imports will outweigh the expansionary effect of exports so that domestic real GDP will decrease.

Negative Net Exports Suppose that net exports are a negative \$5 billion as shown by X_{n2} in Figure 29.4a. This means that our hypothetical economy is importing \$5 billion more of goods than it is exporting. The aggregate expenditures schedule shown as $C + I_g$ in Figure 29.4b therefore overstates the expenditures on domestic output at each level of GDP. We must reduce the sum of expenditures by the \$5 billion net amount spent on imported goods. We do that by subtracting the \$5 billion of net imports from $C + I_g$.

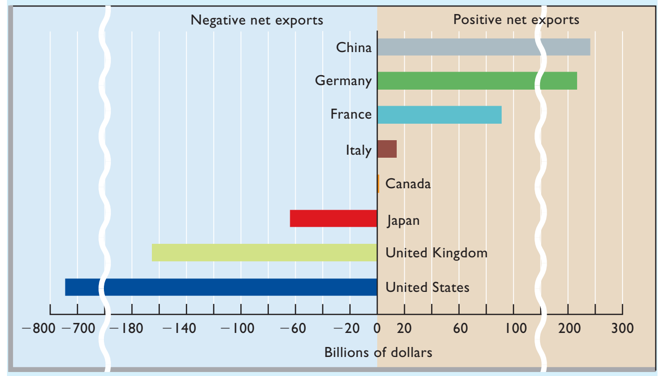
The relevant aggregate expenditures schedule in Figure 29.4b becomes $C + I_g + X_{n2}$ and equilibrium GDP falls from \$470 billion to \$450 billion. Again, a change in net exports of \$5 billion has produced a fourfold change in GDP, reminding us that the multiplier in this example is 4.

This gives us a corollary to our first generalization: Other things equal, negative net exports reduce aggregate expenditures and GDP below what they would be in a closed economy. When imports exceed exports, the contractionary effect of the larger amount of imports outweighs the expansionary effect of the smaller amount of exports, and equilibrium real GDP decreases.

GLOBAL PERSPECTIVE 29.1

Net Exports of Goods, Selected Nations, 2012

Some nations, such as China and Germany, have positive net exports; other countries, such as the United States and the United Kingdom, have negative net exports.



Source: CIA World Factbook, **www.cia.gov**.

Our generalizations of the effects of net exports on GDP mean that a decline in X_n—a decrease in exports or an increase in imports—reduces aggregate expenditures and contracts a nation's GDP. Conversely, an increase in X_n—the result of either an increase in exports or a decrease in imports—increases aggregate expenditures and expands GDP.

As is shown in Global Perspective 29.1, net exports vary greatly among the major industrial nations.

International Economic Linkages

Our analysis of net exports and real GDP suggests how circumstances or policies abroad can affect U.S. GDP.

Prosperity Abroad A rising level of real output and income among U.S. foreign trading partners enables the United States to sell more goods abroad, thus raising U.S. net exports and increasing U.S. real GDP (assuming initially there is excess capacity). There is good reason for Americans to be interested in the prosperity of our trading partners. Their good fortune enables them to buy more of our exports, increasing our income and enabling us in turn to buy more foreign imports. These imported goods are the ultimate benefit of international trade. Prosperity abroad transfers some of that prosperity to Americans.

Exchange Rates Depreciation of the dollar relative to other currencies enables people abroad to obtain more dollars with each unit of their own currencies. The price of U.S. goods in terms of those currencies will fall, stimulating purchases of U.S. exports. Also, U.S. customers will find they need more dollars to buy foreign goods and, consequently, will reduce their spending on imports. If the economy has available capacity, the increased exports and decreased imports will increase U.S. net exports and expand the nation's GDP.

This last example has been cast only in terms of depreciation of the dollar. You should think through the impact that appreciation of the dollar would have on net exports and equilibrium GDP.

A Caution on Tariffs and Devaluations Because higher net exports increase real GDP, countries often look for ways to reduce imports and increase exports during recessions or depressions. Thus, a recession might tempt the U.S. federal government to increase tariffs and devalue the international value of the dollar (say, by supplying massive amounts of dollars in the foreign exchange market) to try to increase net exports. Such an increase of net exports would expand domestic production, reduce domestic unemployment, and help the economy recover.

But this interventionist thinking is too simplistic. Suppose that the United States imposes high tariffs on foreign goods to reduce our imports and thus increase our domestic production and employment. Our imports, however, are our trading partners' exports. So when we restrict our imports to stimulate our economy, we depress the economies of our trading partners. They are likely to retaliate against us by imposing tariffs on our products. If so, our exports to them will decline and our net exports may in fact fall. With retaliation in the picture, it is possible that tariffs may decrease, not increase, our net exports.

That unfortunate possibility became a sad reality during the Great Depression of the 1930s, when various nations, including the United States, imposed trade barriers as a way of reducing domestic unemployment. The result was many rounds of retaliation that simply throttled world trade, worsened the depression, and increased unemployment. Abetting the problem were attempts by some nations to increase their net exports by devaluing their currencies. Other nations simply retaliated by devaluing their own currencies. The international exchange rate system collapsed, and world trade spiraled downward. Economic historians agree that tariffs and devaluations during the 1930s were huge policy mistakes!

Nations are tempted to use tariffs and currency devaluations because, *other things equal*, these policies *do* increase net exports and real GDP. But keep in mind that other things aren't likely to stay equal. In particular, other nations will almost certainly retaliate with their own tariffs and devaluations—the final result being lower net exports and lower GDP for those countries and for our own.

QUICK REVIEW 29.2

- Positive net exports increase aggregate expenditures relative to the closed economy and, other things equal, increase equilibrium GDP.
- Negative net exports decrease aggregate expenditures relative to the closed economy and, other things equal, reduce equilibrium GDP.
- In the open economy, changes in (a) prosperity abroad, (b) tariffs, and (c) exchange rates can affect U.S. net exports and therefore U.S. aggregate expenditures and equilibrium GDP.
- Tariffs and deliberate currency depreciations are unlikely to increase net exports because other nations will retaliate.

Adding the Public Sector

LO29.7 Explain how economists integrate the public sector (government expenditures and taxes) into the aggregate expenditures model.

Our final step in constructing the full aggregate expenditures model is to move the analysis from a private (no-government) open economy to an economy with a public sector (sometimes called a "mixed economy"). This means adding government purchases and taxes to the model.

For simplicity, we will assume that government purchases are independent of the level of GDP and do not alter the consumption and investment schedules. Also, government's net tax revenues—total tax revenues less "negative taxes" in the form of transfer payments—are derived entirely from personal taxes. Finally, a fixed amount of taxes is collected regardless of the level of GDP.

Government Purchases and Equilibrium GDP

Suppose the government decides to purchase $20 billion of goods and services regardless of the level of GDP and tax collections.

Tabular Example Table 29.4 shows the impact of this purchase on the equilibrium GDP. Columns 1 through 4 are carried over from Table 29.2 for the private closed economy, in which the equilibrium GDP was $470 billion. The only new items are exports and imports in column 5 and government purchases in column 6. (Observe in column 5 that net exports are zero.) As shown in column 7, the addition of

TABLE 29.4 The Impact of Government Purchases on Equilibrium GDP

(1) Real Domestic Output and Income (GDP = DI), Billions	(2) Consumption (C), Billions	(3) Savings (S), Billions	(4) Investment (I_g), Billions	(5) Net Exports (X_n), Billions: Exports (X)	(5) Net Exports (X_n), Billions: Imports (M)	(6) Government Purchases (G), Billions	(7) Aggregate Expenditures ($C + I_g + X_n + G$), Billions (2) + (4) + (5) + (6)
(1) $370	$375	$–5	$20	$10	$10	$20	$415
(2) 390	390	0	20	10	10	20	430
(3) 410	405	5	20	10	10	20	445
(4) 430	420	10	20	10	10	20	460
(5) 450	435	15	20	10	10	20	475
(6) 470	450	20	20	10	10	20	490
(7) 490	465	25	20	10	10	20	505
(8) 510	480	30	20	10	10	20	520
(9) 530	495	35	20	10	10	20	535
(10) 550	510	40	20	10	10	20	550

government purchases to private spending ($C + I_g + X_n$) yields a new, higher level of aggregate expenditures ($C + I_g + X_n + G$). Comparing columns 1 and 7, we find that aggregate expenditures and real output are equal at a higher level of GDP. Without government purchases, equilibrium GDP was $470 billion (row 6); *with* government purchases, aggregate expenditures and real output are equal at $550 billion (row 10). Increases in public spending, like increases in private spending, shift the aggregate expenditures schedule upward and produce a higher equilibrium GDP.

Note, too, that government spending is subject to the multiplier. A $20 billion increase in government purchases has increased equilibrium GDP by $80 billion (from $470 billion to $550 billion). The multiplier in this example is 4.

This $20 billion increase in government spending is *not* financed by increased taxes. Shortly, we will demonstrate that increased taxes *reduce* equilibrium GDP.

Graphical Analysis In Figure 29.5, we vertically add $20 billion of government purchases, *G*, to the level of

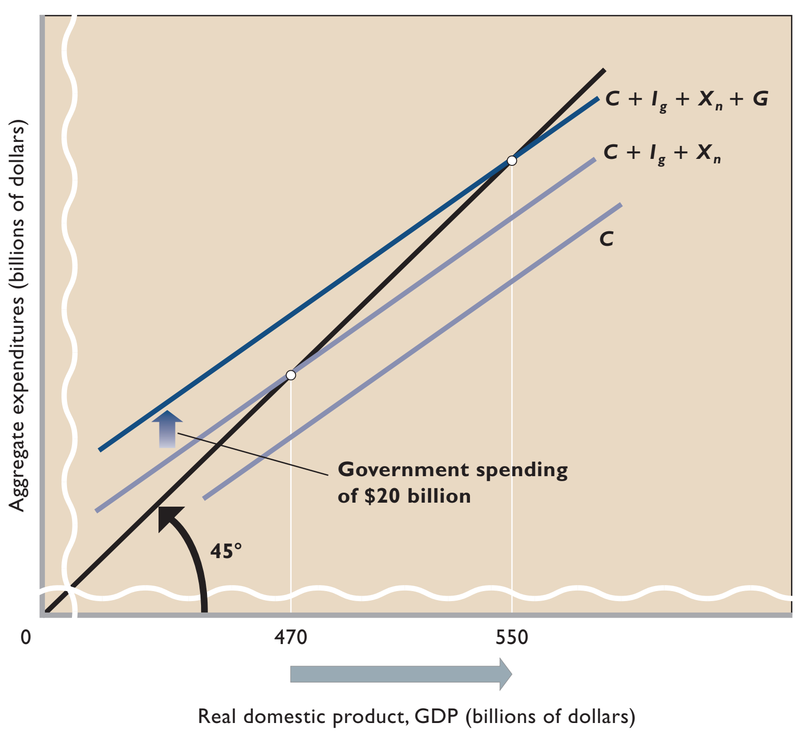


FIGURE 29.5 Government spending and equilibrium GDP. The addition of government expenditures of *G* to our analysis raises the aggregate expenditures ($C + I_g + X_n + G$) schedule and increases the equilibrium level of GDP, as would an increase in *C*, I_g, or X_n.

TABLE 29.5 **Determination of the Equilibrium Levels of Employment, Output, and Income: Private and Public Sectors**

(1) Real Domestic Output and Income (GDP = NI = PI), Billions	(2) Taxes (*T*), Billions	(3) Disposable Income (DI), Billions, (1) − (2)	(4) Consumption (C_a), Billions	(5) Saving (S_a), Billions (3) − (4)	(6) Investment (I_g), Billions	(7) Net Exports (X_n), Billions: Exports (*X*)	Imports (*M*)	(8) Government Purchases (*G*), Billions	(9) Aggregate Expenditures ($C_a + I_g + X_n + G$), Billions, (4) + (6) + (7) + (8)
(1) $370	$20	$350	$360	$−10	$20	$10	$10	$20	$400
(2) 390	20	370	375	−5	20	10	10	20	415
(3) 410	20	390	390	0	20	10	10	20	430
(4) 430	20	410	405	5	20	10	10	20	445
(5) 450	20	430	420	10	20	10	10	20	460
(6) 470	20	450	435	15	20	10	10	20	475
(7) 490	20	470	450	20	20	10	10	20	490
(8) 510	20	490	465	25	20	10	10	20	505
(9) 530	20	510	480	30	20	10	10	20	520
(10) 550	20	530	495	35	20	10	10	20	535

private spending, $C + I_g + X_n$. That added $20 billion raises the aggregate expenditures schedule (private plus public) to $C + I_g + X_n + G$, resulting in an $80 billion increase in equilibrium GDP, from $470 to $550 billion.

A decline in government purchases G will lower the aggregate expenditures schedule in Figure 29.5 and result in a multiplied decline in the equilibrium GDP. Verify in Table 29.4 that if government purchases were to decline from $20 billion to $10 billion, the equilibrium GDP would fall by $40 billion.

Taxation and Equilibrium GDP

The government not only spends but also collects taxes. Suppose it imposes a **lump-sum tax,** which is a tax of a constant amount or, more precisely, a tax yielding the same amount of tax revenue at each level of GDP. Let's assume this tax is $20 billion, so that the government obtains $20 billion of tax revenue at each level of GDP regardless of the level of government purchases.

Tabular Example In Table 29.5, which continues our example, we find taxes in column 2, and we see in column 3 that disposable (after-tax) income is lower than GDP (column 1) by the $20 billion amount of the tax. Because households use disposable income both to consume and to save, the tax lowers both consumption and saving. The MPC and MPS tell us how much consumption and saving will decline as a result of the $20 billion in taxes. Because the MPC is 0.75, the government tax collection of $20 billion will reduce consumption by $15 billion (= 0.75 × $20 billion). Since the MPS is 0.25, saving will drop by $5 billion (= 0.25 × $20 billion).

Columns 4 and 5 in Table 29.5 list the amounts of consumption and saving *at each level of GDP*. Note they are $15 billion and $5 billion smaller than those in Table 29.4. Taxes reduce disposable income relative to GDP by the amount of the taxes. This decline in DI reduces both consumption and saving at each level of GDP. The extent of the C and S reductions depend on the MPC and the MPS.

To find the effect of taxes on equilibrium GDP, we calculate aggregate expenditures again, as shown in column 9, Table 29.5. Aggregate spending is $15 billion less at each level of GDP than it was in Table 29.4. The reason is that after-tax consumption, designated by C_a, is $15 billion less at each level of GDP. A comparison of real output and aggregate expenditures in columns 1 and 9 shows that the aggregate amounts produced and purchased are equal only at $490 billion of GDP (row 7). The $20 billion lump-sum tax has reduced equilibrium GDP by $60 billion, from $550 billion (row 10, Table 29.3) to $490 billion (row 7, Table 29.4).

Graphical Analysis In Figure 29.6 the $20 billion increase in taxes shows up as a $15 (not $20) billion decline in the aggregate expenditures ($C_a + I_g + X_n + G$) schedule. This decline in the schedule results solely from a decline in the consumption C component of aggregate expenditures. The equilibrium GDP falls from $550 billion to $490 billion because of this tax-caused drop in consumption. With no change in government expenditures, tax increases lower the aggregate expenditures schedule relative to the 45° line and reduce the equilibrium GDP.

In contrast to our previous case, a *decrease* in existing taxes will raise the aggregate expenditures schedule in

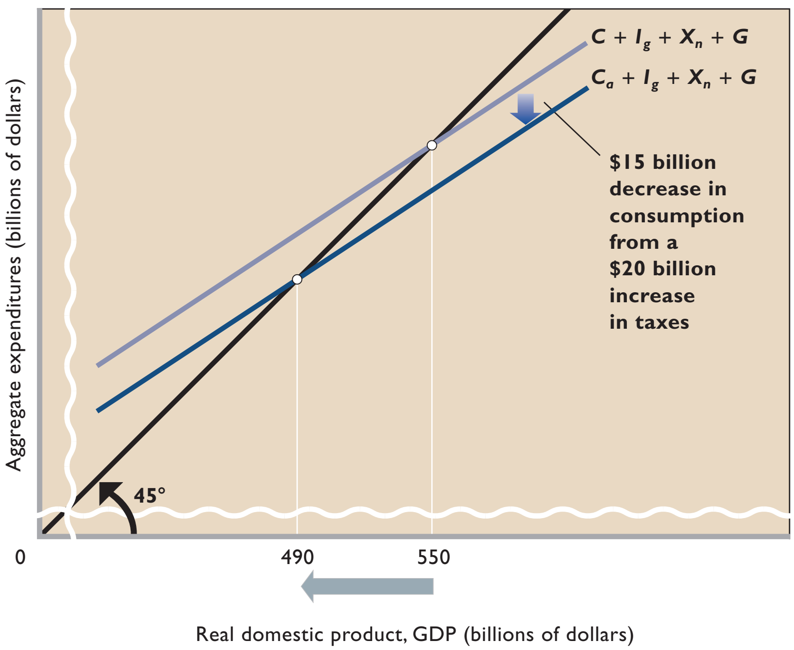


FIGURE 29.6 Taxes and equilibrium GDP. If the MPC is 0.75, the $20 billion of taxes will lower the consumption schedule by $15 billion and cause a $60 billion decline in the equilibrium GDP. In the open economy with government, equilibrium GDP occurs where C_a (after-tax income) $+ I_g + X_n + G =$ GDP. Here that equilibrium is $490 billion.

Figure 29.6 as a result of an increase in consumption at all GDP levels. You should confirm that a tax reduction of $10 billion (from the present $20 billion to $10 billion) would increase the equilibrium GDP from $490 billion to $520 billion.

Differential Impacts You may have noted that equal changes in G and T do not have equivalent impacts on GDP. The $20 billion increase in G in our illustration, subject to the multiplier of 4, produced an $80 billion increase in real GDP. But the $20 billion increase in taxes reduced GDP by only $60 billion. Given an MPC of 0.75, the tax increase of $20 billion reduced consumption by only $15 billion (not $20 billion) because saving also fell by $5 billion. Subjecting the $15 billion decline in consumption to the multiplier of 4, we find the tax increase of $20 billion reduced GDP by $60 billion (not $80 billion).

WORKED PROBLEMS

W29.2
Complete aggregate expenditures model

Table 29.5 and Figure 29.6 constitute the complete aggregate expenditures model for an open economy with government. When total spending equals total production, the economy's output is in equilibrium. In the open mixed economy, equilibrium GDP occurs where

$$C_a + I_g + X_n + G = \text{GDP}$$

Injections, Leakages, and Unplanned Changes in Inventories The related characteristics of equilibrium noted for the private closed economy also apply to the full model. In particular, it is still the case that injections into the income-expenditures stream equal leakages from the income stream. For the private closed economy, $S = I_g$. For the expanded economy, imports and taxes are added leakages. Saving, importing, and paying taxes are all uses of income that subtract from potential consumption. Consumption will now be less than GDP—creating a potential spending gap—in the amount of after-tax saving (S_a), imports (M), and taxes (T). But exports (X) and government purchases (G), along with investment (I_g), are injections into the income-expenditures stream. At the equilibrium GDP, the sum of the leakages equals the sum of injections. In symbols:

$$S_a + M + T = I_g + X + G$$

You should use the data in Table 29.5 to confirm this equality between leakages and injections at the equilibrium GDP of $490 billion. Also, substantiate that a lack of such equality exists at all other possible levels of GDP.

Although not directly shown in Table 29.5, the equilibrium characteristic of "no unplanned changes in inventories" will also be fulfilled at the $490 billion GDP. Because aggregate expenditures equal GDP, all the goods and services produced will be purchased. There will be no unplanned increase in inventories, so firms will have no incentive to reduce their employment and production.

Nor will they experience an unplanned decline in their inventories, which would prompt them to expand their employment and output to replenish their inventories.

Equilibrium versus Full-Employment GDP

LO29.8 Differentiate between equilibrium GDP and full-employment GDP and identify and describe the nature and causes of "recessionary expenditure gaps" and "inflationary expenditure gaps."

A key point about the equilibrium GDP of the aggregate expenditures model is that it need not equal the economy's full-employment GDP. In fact, Keynes specifically designed the model so that it could explain situations like the Great Depression, during which the economy was seemingly stuck at a bad equilibrium in which real GDP was far below potential output. As we will show you in a moment, Keynes also used the model to suggest policy recommendations for moving the economy back toward potential output and full employment.

The fact that equilibrium and potential GDP in the aggregate expenditure model need not match also reveals critical insights about the causes of demand-pull inflation. We will first examine the "expenditure gaps" that give rise to differences between equilibrium and potential GDP and then see how the model helps to explain the recession of 2007–2009 and preview the policies used by the federal government to try to halt and reverse it.

Recessionary Expenditure Gap

Suppose in **Figure 29.7 (Key Graph),** panel (a), that the full-employment level of GDP is $510 billion and the aggregate expenditures schedule is AE_1. (For simplicity, we will now dispense with the $C_a + I_g + X_n + G$ labeling.) This schedule intersects the 45° line to the left of the economy's full-employment output, so the economy's equilibrium GDP of $490 billion is $20 billion short of its full-employment output of $510 billion. According to column 1 in Table 29.2, total employment at the full-employment GDP is 75 million workers. But the economy depicted in Figure 29.7a is employing only 70 million workers; 5 million available workers are not employed. For that reason, the economy is sacrificing $20 billion of output.

A **recessionary expenditure gap** is the amount by which aggregate expenditures *at the full-employment GDP* fall short of those required to achieve the full-employment GDP. Insufficient total spending contracts or depresses the economy. Table 29.5 shows that at the full-employment level of $510 billion (column 1), the corresponding level of aggregate expenditures is only $505 billion (column 9). The recessionary expenditure gap is thus $5 billion, the amount by which the aggregate expenditures curve would have to shift upward to realize equilibrium at the full-employment GDP. Graphically, the recessionary expenditure gap is the *vertical* distance (measured at the full-employment GDP) by which the actual aggregate expenditures schedule AE_1 lies below the hypothetical full-employment aggregate expenditures schedule AE_0. In Figure 29.7a, this recessionary expenditure gap is $5 billion. Because the multiplier is 4, there is a $20 billion differential (the recessionary expenditure gap of $5 billion times the multiplier of 4) between the equilibrium GDP and the full-employment GDP. This $20 billion difference is a negative *GDP gap*—an idea we first developed when discussing cyclical unemployment in Chapter 27.

Keynes's Solution to a Recessionary Expenditure Gap

Keynes pointed to two different policies that a government might pursue to close a recessionary expenditure gap and achieve full employment. The first is to increase government spending. The second is to lower taxes. Both work by increasing aggregate expenditures.

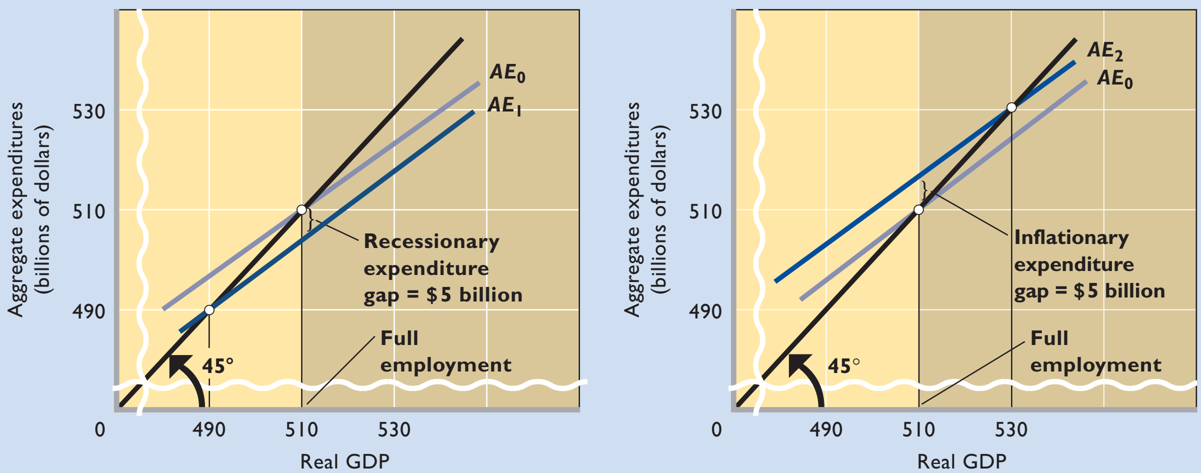
Look back at Figure 29.5. There we showed how an increase in government expenditures G will increase overall aggregate expenditures and, consequently, the equilibrium real GDP. Applying this strategy to the situation in Figure 29.7a, government could completely close the $20 billion negative GDP gap between the initial equilibrium of $490 billion and the economy's potential output of $510 billion if it increased spending by the $5 billion amount of the recessionary expenditure gap. Given the economy's multiplier of 4, the $5 billion increase in G would create a $20 billion increase in equilibrium real GDP, thereby bringing the economy to full employment.

Government also could lower taxes to close the recessionary expenditure gap and thus eliminate the negative GDP gap. Look back at Figure 29.6 in which an increase in taxes resulted in lower after-tax consumption spending and a smaller equilibrium real GDP. Keynes simply suggested a reversal of this process: Since an increase in taxes lowers equilibrium real GDP, a decrease in taxes will raise equilibrium GDP. The decrease in taxes will leave consumers with higher after-tax income. That will lead to higher consumption expenditures and an increase in equilibrium real GDP.

But by how much should the government cut taxes? By exactly $6.67 billion. That is because the MPC is 0.75. The

KEY GRAPH

FIGURE 29.7 Recessionary and inflationary expenditure gaps. The equilibrium and full-employment GDPs may not coincide. (a) A recessionary expenditure gap is the amount by which aggregate expenditures at the full-employment GDP fall short of those needed to achieve the full-employment GDP. Here, the $5 billion recessionary expenditure gap causes a $20 billion negative GDP gap. (b) An inflationary expenditure gap is the amount by which aggregate expenditures at the full-employment GDP exceed those just sufficient to achieve the full-employment GDP. Here, the inflationary expenditure gap is $5 billion; this overspending produces demand-pull inflation.



(a) Recessionary expenditure gap

(b) Inflationary expenditure gap

QUICK QUIZ FOR FIGURE 29.7

1. In the economy depicted:
- **a.** the MPS is 0.50.
- **b.** the MPC is 0.75.
- **c.** the full-employment level of real GDP is $530 billion.
- **d.** nominal GDP always equals real GDP.

2. The inflationary expenditure gap depicted will cause:
- **a.** demand-pull inflation.
- **b.** cost-push inflation.
- **c.** cyclical unemployment.
- **d.** frictional unemployment.

3. The recessionary expenditure gap depicted will cause:
- **a.** demand-pull inflation.
- **b.** cost-push inflation.
- **c.** cyclical unemployment.
- **d.** frictional unemployment.

4. In the economy depicted, the $5 billion inflationary expenditure gap:
- **a.** expands real GDP to $530 billion.
- **b.** leaves real GDP at $510 billion but causes inflation.
- **c.** could be remedied by equal $5 billion increases in taxes and government spending.
- **d.** implies that real GDP exceeds nominal GDP.

Answers: 1. b; 2. a; 3. c; 4. b

tax cut of $6.67 billion will increase consumers' after-tax income by $6.67 billion. They will then increase consumption spending by 0.75 of that amount, or $5 billion. This will increase aggregate expenditures by the $5 billion needed to close the recessionary expenditure gap. The economy's equilibrium real GDP will rise to its potential output of $510 billion.

But a big warning is needed here: As the economy moves closer to its potential output, it becomes harder to justify Keynes's assumption that prices are stuck. As the economy closes its negative GDP gap, nearly all workers are employed and nearly all factories are operating at or near full capacity. In such a situation, there is no massive oversupply of productive resources to keep prices from rising. In fact, economists know from real-world experience that in such situations prices are not fully stuck. Instead, they become increasingly flexible as the economy moves nearer to potential output.

This fact is one of the major limitations of the aggregate expenditures model and is the reason why we will develop a different model that can handle inflation in the next chapter. That being said, it is nevertheless true that the aggregate expenditures model is still very useful despite its inability to handle flexible prices. For instance, as we explained in Chapter 24, even an economy operating near full employment will show sticky or even stuck prices in the short run. In such situations, the intuitions of the aggregate expenditures model will still hold true. The benefit of the aggregate demand–aggregate supply model that we develop in the next chapter is that it also can show us what happens over longer periods, as prices (and wages) become more flexible and are increasingly able to adjust.

WORKED PROBLEMS

W29.3
Expenditure gaps

Inflationary Expenditure Gap

Economists use the term **inflationary expenditure gap** to describe the amount by which an economy's aggregate expenditures *at the full-employment GDP* exceed those just necessary to achieve the full-employment level of GDP. In Figure 29.7b, there is a \$5 billion inflationary expenditure gap at the \$510 billion full-employment GDP. This is shown by the vertical distance between the actual aggregate expenditures schedule AE_2 and the hypothetical schedule AE_0 that would be just sufficient to achieve the \$510 billion full-employment GDP. Thus, the inflationary expenditure gap is the amount by which the aggregate expenditures schedule would have to shift downward to realize equilibrium at the full-employment GDP.

But why does the name "inflationary expenditure gap" contain the word *inflationary*? In particular, what does the situation depicted in Figure 29.7b have to do with inflation? The answer lies in the answer to a different question: *Could the economy actually achieve and maintain an equilibrium real GDP that is substantially above the full-employment output level?*

The unfortunate answer is no. It is unfortunate because if such a thing were possible, then the government could make real GDP as high as it wanted by simply increasing G to an arbitrarily high number. Graphically, it could raise the AE_2 curve in Figure 29.7b as far up as it wanted, thereby increasing equilibrium real GDP as high as it wanted. Living standards would skyrocket! But this is not possible because, by definition, all the available workers in the economy are fully employed at the full-employment output level. Producing slightly more than the full-employment output level for a few months might be possible if you could convince all the workers to work overtime day after day. But there simply is not enough labor to have the economy produce at much more than potential output for any extended period of time.

So what *does* happen in situations in which aggregate expenditures are so high that the model predicts an equilibrium level of GDP beyond potential output? The answer is twofold. First, the economy ends up producing either at potential output or just above potential output due to the limited supply of labor. Second, the economy experiences demand-pull inflation. With the supply of output limited by the supply of labor, high levels of aggregate expenditures simply drive up prices. Nominal GDP will increase because of the higher price level, but real GDP will not.

Application: The Recession of 2007–2009

In December 2007 the U.S. economy entered the longest and one of the deepest recessions since the Great Depression of the 1930s. We will defer discussion of the underlying financial crisis until later chapters, but the ultimate effect of the crisis is easily portrayed through the aggregate expenditures model. We know that the AE_0 line in Figure 29.7a consists of the combined amount of after-tax consumption expenditures (C_a), gross investment expenditures (I_g), net export expenditures (X_n), and government purchases (G) planned at each level of real GDP. During the recession, both after-tax consumption and investment expenditures declined, with planned investment expenditures suffering the largest drop by far.

Aggregate expenditures thus declined, as from AE_0 to AE_1 in Figure 29.7a. This set off a multiple decline in real GDP, illustrated in the figure by the decline from \$510 billion to \$490 billion. In the language of the aggregate expenditures model, a recessionary expenditure gap produced one of the largest negative GDP gaps since the Great Depression. Employment sank by more than 8 million people, and the unemployment rate jumped above 10 percent. As recessions go, this was a big one!

The federal government undertook various Keynesian policies in 2008 and 2009 to try to eliminate the recessionary expenditure gap facing the economy. In 2008, the government provided \$100 billion of tax rebate checks to

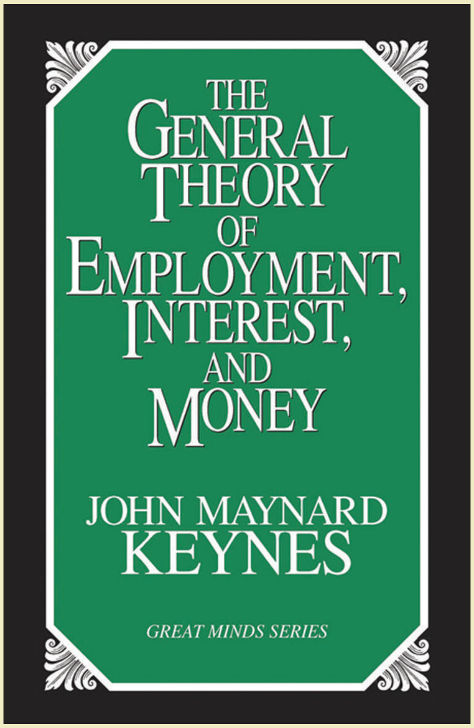
LAST WORD

Say's Law, the Great Depression, and Keynes

The Aggregate Expenditure Theory Emerged as a Critique of Classical Economics and as a Response to the Great Depression.

Until the Great Depression of the 1930s, many prominent economists, including David Ricardo (1772–1823) and John Stuart Mill (1806–1873), believed that the market system would ensure full employment of an economy's resources. These so-called *classical economists* acknowledged that now and then abnormal circumstances such as wars, political upheavals, droughts, speculative crises, and gold rushes would occur, deflecting the economy from full-employment status. But when such deviations occurred, the economy would automatically adjust and soon return to full-employment output. For example, a slump in output and employment would result in lower prices, wages, and interest rates, which in turn would increase consumer spending, employment, and investment spending. Any excess supply of goods and workers would soon be eliminated.

Classical macroeconomists denied that the level of spending in an economy could be too low to bring about the purchase of the entire full-employment output. They based their denial of inadequate spending in part on *Say's law*, attributed to the nineteenth-century French economist J. B. Say (1767–1832). This law is the disarmingly simple idea that the very act of producing goods generates income equal to the value of the goods produced. The production of any output automatically provides the income needed to buy that output. More succinctly stated, *supply creates its own demand*.



Say's law can best be understood in terms of a barter economy. A woodworker, for example, produces or supplies furniture as a means of buying or demanding the food and clothing produced by other workers. The woodworker's supply of furniture is the income that he will "spend" to satisfy his demand for other goods. The goods he buys (demands) will have a total value exactly equal to the goods he produces (supplies). And so it is for other producers and for the entire economy. Demand must be the same as supply!

Assuming that the composition of output is in accord with consumer preferences, all markets would be cleared of their outputs. It would seem that all firms need to do to sell a full-employment output is to produce that level of output. Say's law guarantees there will be sufficient spending to purchase it all.

ORIGIN OF THE IDEA

O29.2

Say's law

The Great Depression of the 1930s called into question the theory that supply creates its own demand (Say's law). In the United States, real GDP declined by 27 percent and the unemployment rate rocketed to nearly 25 percent. Other nations experienced similar impacts. And cyclical unemployment lingered for a decade. An obvious inconsistency exists between a theory that says that unemployment is virtually impossible and the actual occurrence of a 10-year siege of substantial unemployment.

In 1936 British economist John Maynard Keynes (1883–1946) explained why cyclical unemployment could occur in a market economy. In his *General Theory of Employment, Interest, and Money*, Keynes attacked the foundations of classical theory and developed the ideas underlying the aggregate expenditures model. Keynes disputed Say's law, pointing out that not all income need be spent in the same period that it is produced. In fact, some income is always saved. In normal times, that saving is borrowed by businesses to buy capital goods—thereby boosting total spending in the economy. But if expectations about the future grow pessimistic, businesses will slash investment spending and a lot of that saving will not be put to use. The result will be insufficient total spending. Unsold goods will accumulate in producers' warehouses, and producers will respond by reducing their output and discharging workers. A recession or depression will result, and widespread cyclical unemployment will occur. Moreover, said Keynes, recessions or depressions are not likely to correct themselves. In contrast to the more laissez-faire view of the classical economists, Keynes argued that government should play an active role in stabilizing the economy.

taxpayers, hoping that recipients would use most of their checks to buy goods and services. (They didn't! Instead, they used substantial portions of their checks to pay off credit cards and reduce other debt.) In 2009, the federal government enacted a $787 billion stimulus package designed to boost aggregate expenditures, reduce the recessionary expenditure gap, and, through the multiplier effect, increase real GDP and employment.

We will defer discussion and further assessment of these stimulus attempts until Chapter 31, but Figure 29.7a clearly illuminates their purpose. If the government could drive up aggregate expenditures, such as from AE_1 to AE_0, the recession would come to an end and the recovery phase of the business cycle would begin.

QUICK REVIEW 29.3

- Government purchases shift the aggregate expenditures schedule upward and raise equilibrium GDP.
- Taxes reduce disposable income, lower consumption spending and saving, shift the aggregate expenditures schedule downward, and reduce equilibrium GDP.
- A recessionary expenditure gap is the amount by which an economy's aggregate expenditures schedule must shift upward to achieve the full-employment GDP; an inflationary expenditure gap is the amount by which the economy's aggregate expenditures schedule must shift downward to achieve full-employment GDP and eliminate demand-pull inflation.

SUMMARY

LO29.1 Explain how sticky prices relate to the aggregate expenditures model.

The aggregate expenditures model views the total amount of spending in the economy as the primary factor determining the level of real GDP that the economy will produce. The model assumes that the price level is fixed. Keynes made this assumption to reflect the general circumstances of the Great Depression, in which declines in output and employment, rather than declines in prices, were the dominant adjustments made by firms when they faced huge declines in their sales.

LO29.2 Explain how an economy's investment schedule is derived from the investment demand curve and an interest rate.

An investment schedule shows how much investment the firms in an economy are collectively planning to make at each possible level of GDP. In this chapter, we utilize a simple investment schedule in which investment is a constant value and therefore the same at all levels of GDP. That constant value is derived from the investment demand curve by determining what quantity of investment will be demanded at the economy's current real interest rate.

LO29.3 Illustrate how economists combine consumption and investment to depict an aggregate expenditures schedule for a private closed economy and how that schedule can be used to demonstrate the economy's equilibrium level of output (where the total quantity of goods produced equals the total quantity of goods purchased).

For a private closed economy the equilibrium level of GDP occurs when aggregate expenditures and real output are equal or, graphically, where the $C + I_g$ line intersects the 45° line. At any GDP greater than equilibrium GDP, real output will exceed aggregate spending, resulting in unplanned investment in inventories and eventual declines in output and income (GDP). At any below-equilibrium GDP, aggregate expenditures will exceed real output, resulting in unplanned disinvestment in inventories and eventual increases in GDP.

LO29.4 Discuss the two other ways to characterize the equilibrium level of real GDP in a private closed economy: saving = investment, and no unplanned changes in inventories.

At equilibrium GDP, the amount households save (leakages) and the amount businesses plan to invest (injections) are equal. Any excess of saving over planned investment will cause a shortage of total spending, forcing GDP to fall. Any excess of planned investment over saving will cause an excess of total spending, inducing GDP to rise. The change in GDP will in both cases correct the discrepancy between saving and planned investment.

At equilibrium GDP, there are no unplanned changes in inventories. When aggregate expenditures diverge from real GDP, an unplanned change in inventories occurs. Unplanned increases in inventories are followed by a cutback in production and a decline of real GDP. Unplanned decreases in inventories result in an increase in production and a rise of GDP.

Actual investment consists of planned investment plus unplanned changes in inventories and is always equal to saving.

LO29.5 Analyze how changes in equilibrium real GDP can occur in the aggregate expenditures model and describe how those changes relate to the multiplier.

A shift in the investment schedule (caused by changes in expected rates of return or changes in interest rates) shifts the aggregate expenditures curve and causes a new equilibrium level of real GDP. Real GDP changes by more than the amount of the initial change in investment. This multiplier effect ($\Delta GDP/\Delta I_g$) accompanies both increases and decreases in aggregate expenditures and also applies to changes in net exports (X_n) and government purchases (G).

LO29.6 Explain how economists integrate the international sector (exports and imports) into the aggregate expenditures model.

The net export schedule in the model of the open economy relates net exports (exports minus imports) to levels of real GDP. For simplicity, we assume that the level of net exports is the same at all levels of real GDP.

Positive net exports increase aggregate expenditures to a higher level than they would if the economy were "closed" to international trade. Negative net exports decrease aggregate expenditures relative to those in a closed economy, decreasing equilibrium real GDP by a multiple of their amount. Increases in exports or decreases in imports have an expansionary effect on real GDP, while decreases in exports or increases in imports have a contractionary effect.

LO29.7 Explain how economists integrate the public sector (government expenditures and taxes) into the aggregate expenditures model.

Government purchases in the model of the mixed economy shift the aggregate expenditures schedule upward and raise GDP.

Taxation reduces disposable income, lowers consumption and saving, shifts the aggregate expenditures curve downward, and reduces equilibrium GDP.

In the complete aggregate expenditures model, equilibrium GDP occurs where $C_a + I_g + X_n + G = \text{GDP}$. At the equilibrium GDP, *leakages* of after-tax saving (S_a), imports (M), and taxes (T) equal *injections* of investment (I_g), exports (X), and government purchases (G): $S_a + M + T = I_g + X_n + G$. Also, there are no unplanned changes in inventories.

LO29.8 Differentiate between equilibrium GDP and full-employment GDP and identify and describe the nature and causes of "recessionary expenditure gaps" and "inflationary expenditure gaps."

The equilibrium GDP and the full-employment GDP may differ. A recessionary expenditure gap is the amount by which aggregate expenditures at the full-employment GDP fall short of those needed to achieve the full-employment GDP. This gap produces a negative GDP gap (actual GDP minus potential GDP). An inflationary expenditure gap is the amount by which aggregate expenditures at the full-employment GDP exceed those just sufficient to achieve the full-employment GDP. This gap causes demand-pull inflation.

Keynes suggested that the solution to the large negative GDP gap that occurred during the Great Depression was for government to increase aggregate expenditures. It could do this by increasing its own expenditures (G) or by lowering taxes (T) to increase after-tax consumption expenditures (C_a) by households. Because the economy had millions of unemployed workers and massive amounts of unused production capacity, government could boost aggregate expenditures without worrying about creating inflation.

The stuck-price assumption of the aggregate expenditures model is not credible when the economy approaches or attains its full-employment output. With unemployment low and excess production capacity small or nonexistent, an increase in aggregate expenditures will cause inflation along with any increase in real GDP.

TERMS AND CONCEPTS

planned investment	leakage	lump-sum tax
investment schedule	injection	recessionary expenditure gap
aggregate expenditures schedule	unplanned changes in inventories	inflationary expenditure gap
equilibrium GDP	net exports	

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What is an investment schedule and how does it differ from an investment demand curve? **LO29.2**
2. Why does equilibrium real GDP occur where $C + I_g =$ GDP in a private closed economy? What happens to real GDP when $C + I_g$ exceeds GDP? When $C + I_g$ is less than GDP? What two expenditure components of real GDP are purposely excluded in a private closed economy? **LO29.3**
3. Why is saving called a *leakage*? Why is planned investment called an *injection*? Why must saving equal planned investment

at equilibrium GDP in the private closed economy? Are unplanned changes in inventories rising, falling, or constant at equilibrium GDP? Explain. LO29.4

4. Other things equal, what effect will each of the following changes independently have on the equilibrium level of real GDP in the private closed economy? LO29.5
   a. A decline in the real interest rate.
   b. An overall decrease in the expected rate of return on investment.
   c. A sizable, sustained increase in stock prices.
5. Depict graphically the aggregate expenditures model for a private closed economy. Now show a decrease in the aggregate expenditures schedule and explain why the decline in real GDP in your diagram is greater than the decline in the aggregate expenditures schedule. What is the term used for the ratio of a decline in real GDP to the initial drop in aggregate expenditures? LO29.5
6. Assuming the economy is operating below its potential output, what is the impact of an increase in net exports on real GDP? Why is it difficult, if not impossible, for a country to boost its net exports by increasing its tariffs during a global recession? LO29.6
7. What is a recessionary expenditure gap? An inflationary expenditure gap? Which is associated with a positive GDP gap? A negative GDP gap? LO29.8
8. **LAST WORD** What is Say's law? How does it relate to the view held by classical economists that the economy generally will operate at a position on its production possibilities curve (Chapter 1)? Use production possibilities analysis to demonstrate Keynes's view on this matter.

REVIEW QUESTIONS

1. True or False: The aggregate expenditures model assumes flexible prices. LO29.1
2. If total spending is just sufficient to purchase an economy's output, then the economy is: LO29.3
   a. In equilibrium.
   b. In recession.
   c. In debt.
   d. In expansion.
3. True or False: If spending exceeds output, real GDP will decline as firms cut back on production. LO29.3
4. If inventories unexpectedly rise, then production ________ sales and firms will respond by ________ output. LO29.3
   a. Trails; expanding.
   b. Trails; reducing.
   c. Exceeds; expanding.
   d. Exceeds; reducing.
5. If the multiplier is 5 and investment increases by $3 billion, equilibrium real GDP will increase by: LO29.5
   a. $2 billion.
   b. $3 billion.
   c. $8 billion.
   d. $15 billion.
   e. None of the above.
6. A depression abroad will tend to ________ our exports, which in turn will ________ net exports, which in turn will ________ equilibrium real GDP. LO29.6
   a. Reduce; reduce; reduce.
   b. Increase; increase; increase.
   c. Reduce; increase; increase.
   d. Increase; reduce; reduce.
7. Explain graphically the determination of equilibrium GDP for a private economy through the aggregate expenditures model. Now add government purchases (any amount you choose) to your graph, showing its impact on equilibrium GDP. Finally, add taxation (any amount of lump-sum tax that you choose) to your graph and show its effect on equilibrium GDP. Looking at your graph, determine whether equilibrium GDP has increased, decreased, or stayed the same given the sizes of the government purchases and taxes that you selected. LO29.7
8. The economy's current level of equilibrium GDP is $780 billion. The full employment level of GDP is $800 billion. The multiplier is 4. Given those facts, we know that the economy faces ________ expenditure gap of ________. LO29.8
   a. An inflationary; $5 billion.
   b. An inflationary; $10 billion.
   c. An inflationary; $20 billion.
   d. A recessionary; $5 billion.
   e. A recessionary; $10 billion.
   f. A recessionary; $20 billion.
9. If an economy has an inflationary expenditure gap, the government could attempt to bring the economy back toward the full-employment level of GDP by ________ taxes or ________ government expenditures. LO29.8
   a. Increasing; increasing.
   b. Increasing; decreasing.
   c. Decreasing; increasing.
   d. Decreasing; decreasing.

PROBLEMS

1. Assuming the level of investment is $16 billion and independent of the level of total output, complete the following table and determine the equilibrium levels of output and employment in this private closed economy. What are the sizes of the MPC and MPS? LO29.3

Possible Levels of Employment, Millions	Real Domestic Output (GDP = DI), Billions	Consumption, Billions	Saving, Billions
40	$240	$244	$ _____
45	260	260	_____
50	280	276	_____
55	300	292	_____
60	320	308	_____
65	340	324	_____
70	360	340	_____
75	380	356	_____
80	400	372	_____

2. Using the consumption and saving data in problem 1 and assuming investment is $16 billion, what are saving and planned investment at the $380 billion level of domestic output? What are saving and actual investment at that level? What are saving and planned investment at the $300 billion level of domestic output? What are the levels of saving and actual investment? In which direction and by what amount will unplanned investment change as the economy moves from the $380 billion level of GDP to the equilibrium level of real GDP? From the $300 billion level of real GDP to the equilibrium level of GDP? **LO29.4**
3. By how much will GDP change if firms increase their investment by $8 billion and the MPC is 0.80? If the MPC is 0.67? **LO29.5**
4. Suppose that a certain country has an MPC of 0.9 and a real GDP of $400 billion. If its investment spending decreases by $4 billion, what will be its new level of real GDP? **LO29.5**
5. The data in columns 1 and 2 in the table below are for a private closed economy. **LO29.6**
   a. Use columns 1 and 2 to determine the equilibrium GDP for this hypothetical economy.
   b. Now open up this economy to international trade by including the export and import figures of columns 3 and 4. Fill in columns 5 and 6 and determine the equilibrium GDP for the open economy. What is the change in equilibrium GDP caused by the addition of net exports?
   c. Given the original $20 billion level of exports, what would be net exports and the equilibrium GDP if imports were $10 billion greater at each level of GDP?
   d. What is the multiplier in this example?
6. Assume that, without taxes, the consumption schedule of an economy is as follows. **LO29.7**

GDP, Billions	Consumption, Billions
$100	$120
200	200
300	280
400	360
500	440
600	520
700	600

   a. Graph this consumption schedule and determine the MPC.
   b. Assume now that a lump-sum tax is imposed such that the government collects $10 billion in taxes at all levels of GDP. Graph the resulting consumption schedule and compare the MPC and the multiplier with those of the pretax consumption schedule.
7. Refer to columns 1 and 6 in the table for problem 5. Incorporate government into the table by assuming that it plans to tax and spend $20 billion at each possible level of GDP. Also assume that the tax is a personal tax and that government spending does not induce a shift in the private aggregate expenditures schedule. What is the change in equilibrium GDP caused by the addition of government? **LO29.7**
8. **ADVANCED ANALYSIS** Assume that the consumption schedule for a private open economy is such that consumption $C = 50 + 0.8Y$. Assume further that planned investment I_g and net exports X_n are independent of the level of real GDP and constant at $I_g = 30$ and $X_n = 10$. Recall also that, in

(1) Real Domestic Output (GDP = DI), Billions	(2) Aggregate Expenditures, Private Closed Economy, Billions	(3) Exports, Billions	(4) Imports, Billions	(5) Net Exports, Billions	(6) Aggregate Expenditures, Private Open Economy, Billions
$200	$240	$20	$30	$ _____	$ _____
250	280	20	30	_____	_____
300	320	20	30	_____	_____
350	360	20	30	_____	_____
400	400	20	30	_____	_____
450	440	20	30	_____	_____
500	480	20	30	_____	_____
550	520	20	30	_____	_____

equilibrium, the real output produced (Y) is equal to aggregate expenditures: $Y = C + I_g + X_n$. **LO29.7**

a. Calculate the equilibrium level of income or real GDP for this economy.

b. What happens to equilibrium Y if I_g changes to 10? What does this outcome reveal about the size of the multiplier?

9. Refer to the accompanying table in answering the questions that follow: **LO29.8**

(1) Possible Levels of Employment, Millions	(2) Real Domestic Output, Billions	(3) Aggregate Expenditures ($C_a + I_g + X_n + G$), Billions
90	$500	$520
100	550	560
110	600	600
120	650	640
130	700	680

a. If full employment in this economy is 130 million, will there be an inflationary expenditure gap or a recessionary expenditure gap? What will be the consequence of this gap? By how much would aggregate expenditures in column 3 have to change at each level of GDP to eliminate the inflationary expenditure gap or the recessionary expenditure gap? What is the multiplier in this example?

b. Will there be an inflationary expenditure gap or a recessionary expenditure gap if the full-employment level of output is $500 billion? By how much would aggregate expenditures in column 3 have to change at each level of GDP to eliminate the gap? What is the multiplier in this example?

c. Assuming that investment, net exports, and government expenditures do not change with changes in real GDP, what are the sizes of the MPC, the MPS, and the multiplier?

10. Answer the following questions, which relate to the aggregate expenditures model: **LO29.8**

a. If C_a is $100, I_g is $50, X_n is −$10, and G is $30, what is the economy's equilibrium GDP?

b. If real GDP in an economy is currently $200, C_a is $100, I_g is $50, X_n is −$10, and G is $30, will the economy's real GDP rise, fall, or stay the same?

c. Suppose that full-employment (and full-capacity) output in an economy is $200. If C_a is $150, I_g is $50, X_n is −$10, and G is $30, what will be the macroeconomic result?

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **30**

Aggregate Demand and Aggregate Supply

Learning Objectives

LO30.1 Define aggregate demand (AD) and explain how its downward slope is the result of the real-balances effect, the interest-rate effect, and the foreign purchases effect.

LO30.2 Explain the factors that cause changes (shifts) in AD.

LO30.3 Define aggregate supply (AS) and explain how it differs in the immediate short run, the short run, and the long run.

LO30.4 Explain the factors that cause changes (shifts) in AS.

LO30.5 Discuss how AD and AS determine an economy's equilibrium price level and level of real GDP.

LO30.6 Describe how the AD-AS model explains periods of demand-pull inflation, cost-push inflation, and recession.

LO30.7 (Appendix) Identify how the aggregate demand curve relates to the aggregate expenditures model.

During the recession of 2007–2009, the economic terms *aggregate demand* and *aggregate supply* moved from the obscurity of economic journals and textbooks to the spotlight of national newspapers, Web sites, radio, and television.

The media and public asked: Why had *aggregate demand* declined, producing the deepest recession and highest rate of unemployment since 1982? Why hadn't the reductions in interest rates by the Federal Reserve boosted *aggregate demand*? Would the federal government's $787 billion stimulus package increase *aggregate demand* and reduce unemployment, as intended? Would a resurgence of oil prices and other energy prices reduce *aggregate supply,* choking off an economic expansion?

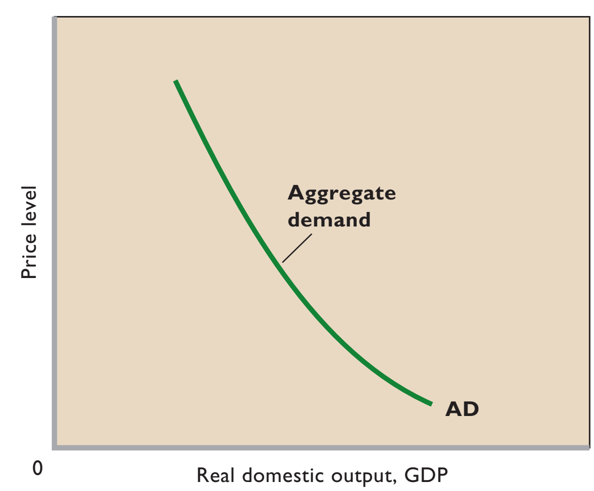
Aggregate demand and aggregate supply are the featured elements of the **aggregate demand–aggregate supply model (AD-AS model),** the focus of this chapter. The aggregate expenditures model of the previous chapter is an immediate-short-run model, in which prices are assumed to be fixed. In contrast, the AD-AS model in this chapter is a "variable price–variable output" model that allows both the price level and level of real GDP to change. It can also show longer time horizons, distinguishing between the immediate short run, the short run, and the long run. Further, in subsequent chapters, we will see that the AD-AS model easily depicts fiscal and monetary policies such as those used in 2008 and 2009 to try to halt the downward slide of the economy and promote its recovery.

Aggregate Demand

LO30.1 Define aggregate demand (AD) and explain how its downward slope is the result of the real-balances effect, the interest-rate effect, and the foreign purchases effect.

Aggregate demand is a schedule or curve that shows the amount of a nation's output (real GDP) that buyers collectively desire to purchase at each possible price level. These buyers include the nation's households, businesses, and government along with consumers located abroad (households, businesses, and governments in other nations). The relationship between the price level (as measured by the GDP price index) and the amount of real GDP demanded is inverse or negative: When the price level rises, the quantity of real GDP demanded decreases; when the price level falls, the quantity of real GDP demanded increases.

FIGURE 30.1 The aggregate demand curve. The downsloping aggregate demand curve AD indicates an inverse (or negative) relationship between the price level and the amount of real output purchased.



Aggregate Demand Curve

The inverse relationship between the price level and real GDP is shown in Figure 30.1, where the aggregate demand curve AD slopes downward, as does the demand curve for an individual product.

Why the downward slope? The explanation is *not* the same as that for why the demand for a single product slopes downward. That explanation centered on the income effect and the substitution effect. When the price of an *individual* product falls, the consumer's (constant) nominal income allows a larger purchase of the product (the income effect). And, as price falls, the consumer wants to buy more of the product because it becomes relatively less expensive than other goods (the substitution effect).

But these explanations do not work for aggregates. In Figure 30.1, when the economy moves down its aggregate demand curve, it moves to a lower general price level. But our circular flow model tells us that when consumers pay lower prices for goods and services, less nominal income flows to resource suppliers in the form of wages, rents, interest, and profits. As a result, a decline in the price level does not necessarily mean an increase in the nominal income of the economy as a whole. Thus, a decline in the price level need not produce an income effect, where more output is purchased because lower nominal prices leave buyers with greater real income.

Similarly, in Figure 30.1, prices in general are falling as we move down the aggregate demand curve, so the rationale for the substitution effect (where more of a specific product is purchased because it becomes cheaper relative to all other products) is not applicable. There is no *overall* substitution effect among domestically produced goods when the price level falls.

If the conventional substitution and income effects do not explain the downward slope of the aggregate demand curve, what does? The explanation rests on three effects of a price-level change.

Real-Balances Effect A change in the price level produces a **real-balances effect.** Here is how it works: A higher price level reduces the real value or purchasing power of the public's accumulated savings balances. In particular, the real value of assets with fixed money values, such as savings accounts or bonds, diminishes. Because a higher price level erodes the purchasing power of such assets, the public is poorer in real terms and will reduce its spending. A household might buy a new car or a plasma TV if the purchasing power of its financial asset balances is, say, \$50,000. But if inflation erodes the purchasing power of its asset balances to \$30,000, the household may defer its purchase. So a higher price level means less consumption spending.

ORIGIN OF THE IDEA

O30.1
Real-balances effect

Interest-Rate Effect The aggregate demand curve also slopes downward because of the **interest-rate effect.** When we draw an aggregate demand curve, we assume that the supply of money in the economy is fixed. But when the price level rises, consumers need more money for purchases and businesses need more money to meet their payrolls and to buy other resources. A \$10 bill will do when the price of an item is \$10, but a \$10 bill plus a \$1 bill is needed when the item costs \$11. In short, a higher price level increases the demand for money. So, given a fixed supply of money, an increase in money demand will drive up the price paid for its use. That price is the interest rate.

Higher interest rates curtail investment spending and interest-sensitive consumption spending. Firms that expect a 6 percent rate of return on a potential purchase of capital will find that investment potentially profitable when the interest rate is, say, 5 percent. But the investment will be unprofitable and will not be made when the interest rate has risen to 7 percent. Similarly, consumers may decide not to purchase a new house or new automobile when the interest rate on loans goes up. So, by increasing the demand for money and consequently the interest rate, a higher price level reduces the amount of real output demanded.

Foreign Purchases Effect The final reason why the aggregate demand curve slopes downward is the **foreign purchases effect.** When the U.S. price level rises relative to foreign price levels (and exchange rates do not respond quickly or completely), foreigners buy fewer U.S. goods and Americans buy more foreign goods. Therefore, U.S. exports fall and U.S. imports rise. In short, the rise in the price level reduces the quantity of U.S. goods demanded as net exports.

These three effects, of course, work in the opposite direction for a decline in the price level. A decline in the price level increases consumption through the real-balances effect and interest-rate effect; increases investment through the interest-rate effect; and raises net exports by increasing exports and decreasing imports through the foreign purchases effect.

Changes in Aggregate Demand

LO30.2 Explain the factors that cause changes (shifts) in AD.

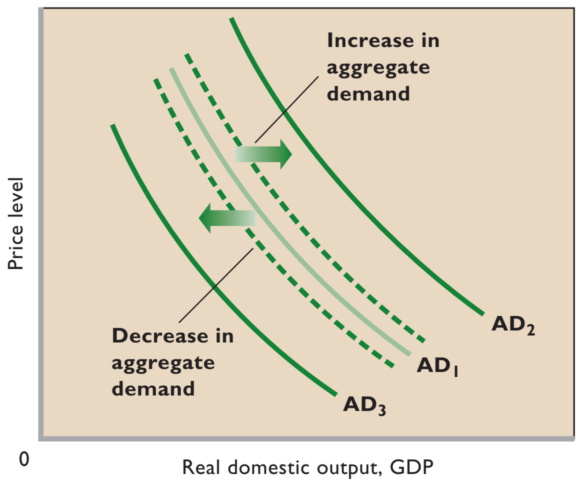
Other things equal, a change in the price level will change the amount of aggregate spending and therefore change the amount of real GDP demanded by the economy. Movements along a fixed aggregate demand curve represent these changes in real GDP. However, if one or more of those "other things" change, the entire aggregate demand curve will shift. We call these other things **determinants of aggregate demand** or, less formally, *aggregate demand shifters.* They are listed in Figure 30.2.

Changes in aggregate demand involve two components:

- A change in one of the determinants of aggregate demand that directly changes the amount of real GDP demanded.
- A multiplier effect that produces a greater ultimate change in aggregate demand than the initiating change in spending.

In Figure 30.2, the full rightward shift of the curve from AD_1 to AD_2 shows an increase in aggregate demand, separated into these two components. The horizontal distance between AD_1 and the broken curve to its right illustrates an initial increase in spending, say, \$5 billion of added investment. If the economy's MPC is 0.75, for example, then the simple multiplier is 4. So the aggregate demand curve shifts rightward from AD_1 to AD_2—four times the distance between AD_1 and the broken line. The multiplier process magnifies the initial change in spending into successive

FIGURE 30.2 Changes in aggregate demand. A change in one or more of the listed determinants of aggregate demand will shift the aggregate demand curve. The rightward shift from AD_1 to AD_2 represents an increase in aggregate demand; the leftward shift from AD_1 to AD_3 shows a decrease in aggregate demand. The vertical distances between AD_1 and the dashed lines represent the initial changes in spending. Through the multiplier effect, that spending produces the full shifts of the curves.



Determinants of Aggregate Demand: Factors That Shift the Aggregate Demand Curve

1. Change in consumer spending
   a. Consumer wealth
   b. Consumer expectations
   c. Household borrowing
   d. Taxes
2. Change in investment spending
   a. Interest rates
   b. Expected returns
      - Expected future business conditions
      - Technology
      - Degree of excess capacity
      - Business taxes
3. Change in government spending
4. Change in net export spending
   a. National income abroad
   b. Exchange rates

rounds of new consumption spending. After the shift, \$20 billion (= \$5 × 4) of additional real goods and services are demanded at each price level.

Similarly, the leftward shift of the curve from AD_1 to AD_3 shows a decrease in aggregate demand, the lesser amount of real GDP demanded at each price level. It also involves the initial decline in spending (shown as the horizontal distance between AD_1 and the dashed line to its left), followed by multiplied declines in consumption spending and the ultimate leftward shift to AD_3.

Let's examine each of the determinants of aggregate demand listed in Figure 30.2.

Consumer Spending

Even when the U.S. price level is constant, domestic consumers may alter their purchases of U.S.-produced real output. If those consumers decide to buy more output at each price level, the aggregate demand curve will shift to the right, as from AD_1 to AD_2 in Figure 30.2. If they decide to buy less output, the aggregate demand curve will shift to the left, as from AD_1 to AD_3.

Several factors other than a change in the price level may change consumer spending and therefore shift the aggregate demand curve. As Figure 30.2 shows, those factors are real consumer wealth, consumer expectations, household debt, and taxes. Because our discussion here parallels that of Chapter 28, we will be brief.

Consumer Wealth Consumer wealth is the total dollar value of all assets owned by consumers in the economy less the dollar value of their liabilities (debts). Assets include stocks, bonds, and real estate. Liabilities include mortgages, car loans, and credit card balances.

Consumer wealth sometimes changes suddenly and unexpectedly due to surprising changes in asset values. An unforeseen increase in the stock market is a good example. The increase in wealth prompts pleasantly surprised consumers to save less and buy more out of their current incomes than they had previously been planning. The resulting increase in consumer spending—the so-called *wealth effect*—shifts the aggregate demand curve to the right. In contrast, an unexpected decline in asset values will cause an unanticipated reduction in consumer wealth at each price level. As consumers tighten their belts in response to the bad news, a "reverse wealth effect" sets in. Unpleasantly surprised consumers increase savings and reduce consumption, thereby shifting the aggregate demand curve to the left.

Household Borrowing Consumers can increase their consumption spending by borrowing. Doing so shifts the aggregate demand curve to the right. By contrast, a decrease in borrowing for consumption purposes shifts the aggregate demand curve to the left. The aggregate demand curve will also shift to the left if consumers increase their savings rates to pay off their debts. With more money flowing to debt repayment, consumption expenditures decline and the AD curve shifts left.

Consumer Expectations Changes in expectations about the future may alter consumer spending. When people expect their future real incomes to rise, they tend to spend more of their current incomes. Thus, current consumption spending increases (current saving falls) and the aggregate demand curve shifts to the right. Similarly, a widely held expectation of surging inflation in the near future may increase aggregate demand today because consumers will want to buy products before their prices escalate. Conversely, expectations of lower future income or lower future prices may reduce current consumption and shift the aggregate demand curve to the left.

Personal Taxes A reduction in personal income tax rates raises take-home income and increases consumer purchases at each possible price level. Tax cuts shift the aggregate demand curve to the right. Tax increases reduce consumption spending and shift the curve to the left.

Investment Spending

Investment spending (the purchase of capital goods) is a second major determinant of aggregate demand. A decline in investment spending at each price level will shift the aggregate demand curve to the left. An increase in investment spending will shift it to the right. In Chapter 28 we saw that investment spending depends on the real interest rate and the expected return from investment.

Real Interest Rates Other things equal, an increase in real interest rates will raise borrowing costs, lower investment spending, and reduce aggregate demand. We are not referring here to the "interest-rate effect" that results from a change in the price level. Instead, we are identifying a change in the real interest rate resulting from, say, a change in a nation's money supply. An increase in the money supply lowers the interest rate, thereby increasing investment and aggregate demand. A decrease in the money supply raises the interest rate, reducing investment and decreasing aggregate demand.

Expected Returns Higher expected returns on investment projects will increase the demand for capital goods and shift the aggregate demand curve to the right. Alternatively, declines in expected returns will decrease investment and shift the curve to the left. Expected returns, in turn, are influenced by several factors:

- ***Expectations about future business conditions*** If firms are optimistic about future business conditions, they are more likely to forecast high rates of return on current investment and therefore may invest more today. On the other hand, if they think the economy will deteriorate in the future, they will forecast low rates of return and perhaps will invest less today.
- ***Technology*** New and improved technologies enhance expected returns on investment and thus increase aggregate demand. For example, recent advances in microbiology have motivated pharmaceutical companies to establish new labs and production facilities.
- ***Degree of excess capacity*** A rise in excess capacity—unused capital—will reduce the expected return on new investment and hence decrease aggregate demand. Other things equal, firms operating factories at well below capacity have little incentive to build new factories. But when firms discover that their excess capacity is dwindling or has completely disappeared, their expected returns on new investment in factories and capital equipment rise. Thus, they increase their investment spending, and the aggregate demand curve shifts to the right.
- ***Business taxes*** An increase in business taxes will reduce after-tax profits from capital investment and lower expected returns. So investment and aggregate demand will decline. A decrease in business taxes will have the opposite effects.

The variability of interest rates and expected returns makes investment highly volatile. In contrast to consumption, investment spending rises and falls often, independent of changes in total income. Investment, in fact, is the least stable component of aggregate demand.

Government Spending

Government purchases are the third determinant of aggregate demand. An increase in government purchases (for example, more transportation projects) will shift the aggregate demand curve to the right, as long as tax collections and interest rates do not change as a result. In contrast, a reduction in government spending (for example, less military equipment) will shift the curve to the left.

Net Export Spending

The final determinant of aggregate demand is net export spending. Other things equal, higher U.S. *exports* mean an increased foreign demand for U.S. goods. So a rise in net exports (higher exports relative to imports) shifts the aggregate demand curve to the right. In contrast, a decrease in U.S. net exports shifts the aggregate demand curve leftward. (These changes in net exports are *not* those prompted by a change in the U.S. price level—those associated with the foreign purchases effect. The changes here are shifts of the AD curve, not movements along the AD curve.)

What might cause net exports to change, other than the price level? Two possibilities are changes in national income abroad and changes in exchange rates.

National Income Abroad Rising national income abroad encourages foreigners to buy more products, some of which are made in the United States. U.S. net exports thus rise, and the U.S. aggregate demand curve shifts to the right. Declines in national income abroad do the opposite: They reduce U.S. net exports and shift the U.S. aggregate demand curve to the left.

Exchange Rates Changes in the dollar's exchange rate—the price of foreign currencies in terms of the U.S. dollar—may affect U.S. exports and therefore aggregate demand. Suppose the dollar depreciates in terms of the euro (meaning the euro appreciates in terms of the dollar). The new, relatively lower value of dollars and higher value of euros enables European consumers to obtain more dollars with each euro. From their perspective, U.S. goods are now less expensive; it takes fewer euros to obtain them. So European consumers buy more U.S. goods, and U.S. exports rise. But American consumers can now obtain fewer euros for each dollar. Because they must pay more dollars to buy European goods, Americans reduce their imports. U.S. exports rise and U.S. imports fall. Conclusion: Dollar depreciation increases net exports (imports go down; exports go up) and therefore increases aggregate demand.

Dollar appreciation has the opposite effects: Net exports fall (imports go up; exports go down) and aggregate demand declines.

QUICK REVIEW 30.1

- Aggregate demand reflects an inverse relationship between the price level and the amount of real output demanded.
- Changes in the price level create real-balances, interest-rate, and foreign purchases effects that explain the downward slope of the aggregate demand curve.
- Changes in one or more of the determinants of aggregate demand (Figure 30.2) alter the amounts of real GDP demanded at each price level; they shift the aggregate demand curve. The multiplier effect magnifies initial changes in spending into larger changes in aggregate demand.
- An increase in aggregate demand is shown as a rightward shift of the aggregate demand curve; a decrease, as a leftward shift of the curve.

Aggregate Supply

LO30.3 Define aggregate supply (AS) and explain how it differs in the immediate short run, the short run, and the long run.

Aggregate supply is a schedule or curve showing the relationship between a nation's price level and the amount of real domestic output that firms in the economy produce. This relationship varies depending on the time horizon and how quickly output prices and input prices can change. We will define three time horizons:

- In the *immediate short run*, both input prices as well as output prices are fixed.
- In the *short run*, input prices are fixed, but output prices can vary.
- In the *long run*, input prices as well as output prices can vary.

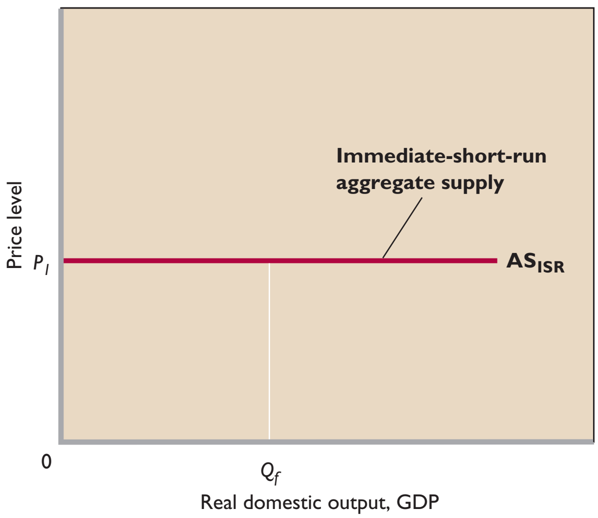
In Chapter 24, we discussed both the immediate short run and the long run in terms of how an automobile maker named Buzzer Auto responds to changes in the demand for its new car, the Prion. Here we extend the logic of that chapter to the economy as a whole to discuss how total output varies with the price level in the immediate short run, the short run, and the long run. As you will see, the relationship between the price level and total output is different in each of the three time horizons because input prices are stickier than output prices. Although both sets of prices become more flexible as time passes, output prices usually adjust more rapidly.

Aggregate Supply in the Immediate Short Run

Depending on the type of firm, the immediate short run can last anywhere from a few days to a few months. It lasts as long as *both* input prices and output prices stay fixed. Input prices are fixed in both the immediate short run and the short run by contractual agreements. In particular, 75 percent of the average firm's costs are wages and salaries—and these are almost always fixed by labor contracts for months or years at a time. As a result, they are usually fixed for a much longer duration than output prices, which can begin to change within a few days or a few months depending on the type of firm.

That being said, output prices are also typically fixed in the immediate short run. This is most often caused by firms setting fixed prices for their customers and then agreeing to supply whatever quantity demanded results at those fixed prices. For instance, once an appliance manufacturer sets its annual list prices for refrigerators, stoves,

FIGURE 30.3 Aggregate supply in the immediate short run. In the immediate short run, the aggregate supply curve AS_{ISR} is horizontal at the economy's current price level, P_1. With output prices fixed, firms collectively supply the level of output that is demanded at those prices.



ovens, and microwaves, it is obligated to supply however many or few appliances customers want to buy at those prices. Similarly, a catalogue company is obliged to sell however much customers want to buy of its products at the prices listed in its current catalogue. And it is obligated to supply those quantities demanded until it sends out its next catalogue.

With output prices fixed and firms selling however much customers want to purchase at those fixed prices, the **immediate-short-run aggregate supply curve** AS_{ISR} is a horizontal line, as shown in Figure 30.3. The AS_{ISR} curve is horizontal at the overall price level P_1, which is calculated from all of the individual prices set by the various firms in the economy. Its horizontal shape implies that the total amount of output supplied in the economy depends directly on the volume of spending that results at price level P_1. If total spending is low at price level P_1, firms will supply a small amount of output to match the low level of spending. If total spending is high at price level P_1, they will supply a high level of output to match the high level of spending. The amount of output that results may be higher than or lower than the economy's full-employment output level Q_f.

Notice, however, that firms will respond in this manner to changes in total spending only as long as output prices remain fixed. As soon as firms are able to change their product prices, they can respond to changes in aggregate spending not only by increasing or decreasing output but also by raising or lowering prices. This is the situation that leads to the upsloping short-run aggregate supply curve that we discuss next.

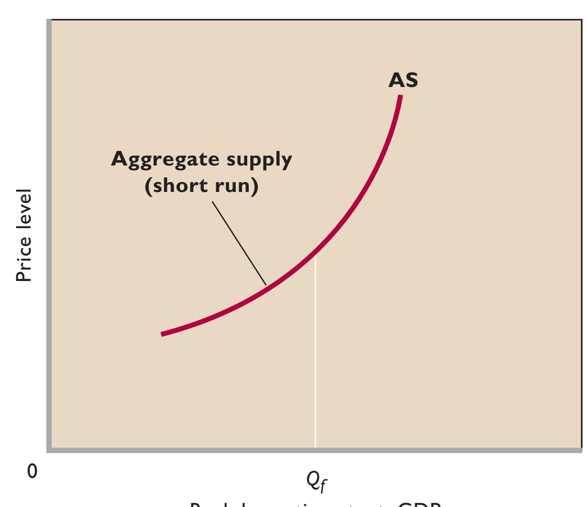
Aggregate Supply in the Short Run

The short run begins after the immediate short run ends. As it relates to macroeconomics, the short run is a period of time during which output prices are flexible, but input prices are either totally fixed or highly inflexible.

These assumptions about output prices and input prices are general—they relate to the economy in the aggregate. Naturally, some input prices are more flexible than others. Since gasoline prices are quite flexible, a package delivery firm like UPS that uses gasoline as an input will have at least one very flexible input price. On the other hand, wages at UPS are set by five-year labor contracts negotiated with its drivers' union, the Teamsters. Because wages are the firm's largest and most important input cost, UPS faces overall input prices that are inflexible for several years at a time. Thus, its "short run"—the period when it can change its shipping prices but not its substantially fixed input prices—is actually quite long. Keep this example in mind as we derive the short-run aggregate supply for the entire economy. Its applicability does not depend on some arbitrary definition of how long the "short run" should be. Instead, the short-run for which the model is relevant is any period of time during which output prices are flexible, but input prices are fixed or nearly fixed.

As illustrated in Figure 30.4, the **short-run aggregate supply curve** AS slopes upward because, with input prices

FIGURE 30.4 The aggregate supply curve (short run). The upsloping aggregate supply curve AS indicates a direct (or positive) relationship between the price level and the amount of real output that firms will offer for sale. The AS curve is relatively flat below the full-employment output because unemployed resources and unused capacity allow firms to respond to price-level rises with large increases in real output. It is relatively steep beyond the full-employment output because resource shortages and capacity limitations make it difficult to expand real output as the price level rises.



fixed, changes in the price level will raise or lower real firm profits. To see how this works, consider an economy that has only a single multiproduct firm called Mega Buzzer and in which the firm's owners must receive a real profit of \$20 to produce the full-employment output of 100 units. Assume the owner's only input (aside from entrepreneurial talent) is 10 units of hired labor at \$8 per worker, for a total wage cost of \$80. Also, assume that the 100 units of output sell for \$1 per unit, so total revenue is \$100. Mega Buzzer's nominal profit is \$20 (= \$100 − \$80), and using the \$1 price to designate the base-price index of 100, its real profit is also \$20 (= \$20/1.00). Well and good; the full-employment output is produced.

Next consider what will happen if the price of Mega Buzzer's output doubles. The doubling of the price level will boost total revenue from \$100 to \$200, but since we are discussing the short run during which input prices are fixed, the \$8 nominal wage for each of the 10 workers will remain unchanged so that total costs stay at \$80. Nominal profit will rise from \$20 (= \$100 − \$80) to \$120 (= \$200 − \$80). Dividing that \$120 profit by the new price index of 200 (= 2.0 in hundredths), we find that Mega Buzzer's real profit is now \$60. The rise in the real reward from \$20 to \$60 prompts the firm (economy) to produce more output. Conversely, price-level declines reduce real profits and cause the firm (economy) to reduce its output. So, in the short run, there is a direct, or positive, relationship between the price level and real output. When the price level rises, real output rises and when the price level falls, real output falls. The result is an upsloping short-run aggregate supply curve.

Notice, however, that the upslope of the short-run aggregate supply curve is not constant. It is flatter at outputs below the full-employment output level Q_f and steeper at outputs above it. This has to do with the fact that per-unit production costs underlie the short-run aggregate supply curve. Recall from Chapter 27 that

$$\text{Per-unit production cost} = \frac{\text{total input cost}}{\text{units of output}}$$

The per-unit production cost of any specific level of output establishes that output's price level because the associated price level must cover all the costs of production, including profit "costs."

As the economy expands in the short run, per-unit production costs generally rise because of reduced efficiency. But the extent of that rise depends on where the economy is operating relative to its capacity. When the economy is operating below its full-employment output, it has large amounts of unused machinery and equipment and large numbers of unemployed workers. Firms can put these idle human and property resources back to work with little upward pressure on per-unit production costs. And as output expands, few if any shortages of inputs or production bottlenecks will arise to raise per-unit production costs. That is why the slope of the short-run aggregate supply curve increases only slowly at output levels below the full-employment output level Q_f.

On the other hand, when the economy is operating beyond Q_f, the vast majority of its available resources are already employed. Adding more workers to a relatively fixed number of highly used capital resources such as plant and equipment creates congestion in the workplace and reduces the efficiency (on average) of workers. Adding more capital, given the limited number of available workers, leaves equipment idle and reduces the efficiency of capital. Adding more land resources when capital and labor are highly constrained reduces the efficiency of land resources. Under these circumstances, total input costs rise more rapidly than total output. The result is rapidly rising per-unit production costs that give the short-run aggregate supply curve its rapidly increasing slope at output levels beyond Q_f.

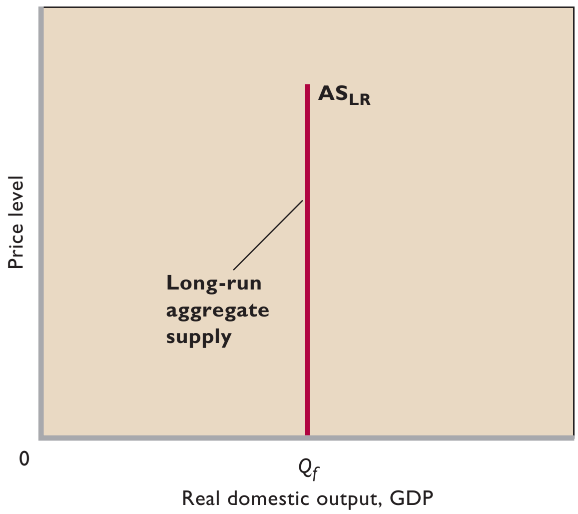
Aggregate Supply in the Long Run

In macroeconomics, the long run is the time horizon over which both input prices as well as output prices are flexible. It begins after the short run ends. Depending on the type of firm and industry, this may be from a couple of weeks to several years in the future. But for the economy as a whole, it is the time horizon over which all output and input prices—including wage rates—are fully flexible.

The **long-run aggregate supply curve** AS_{LR} is vertical at the economy's full-employment output Q_f, as shown in Figure 30.5. The vertical curve means that in the long run the economy will produce the full-employment output level no matter what the price level is. How can this be? Shouldn't higher prices cause firms to increase output? The explanation lies in the fact that in the long run when both input prices as well as output prices are flexible, profit levels will always adjust to give firms exactly the right profit incentive to produce exactly the full-employment output level, Q_f.

To see why this is true, look back at the short-run aggregate supply curve AS shown in Figure 30.4. Suppose that the economy starts out producing at the full-employment output level Q_f and that the price level at that moment has an index value of $P = 100$. Now suppose that output prices double, so that the price index goes to $P = 200$. We previously demonstrated for our single-firm economy that this doubling of the price level would cause profits to rise in

FIGURE 30.5 Aggregate supply in the long run. The long-run aggregate supply curve AS_{LR} is vertical at the full-employment level of real GDP (Q_f) because in the long run wages and other input prices rise and fall to match changes in the price level. So price-level changes do not affect firms' profits and thus they create no incentive for firms to alter their output.



the short run and that the higher profits would motivate the firm to increase output.

This outcome, however, is totally dependent on the fact that input prices are fixed in the short run. Consider what will happen in the long run when they are free to change. Firms can only produce beyond the full-employment output level by running factories and businesses at extremely high rates. This creates a great deal of demand for the economy's limited supply of productive resources. In particular, labor is in great demand because the only way to produce beyond full employment is if workers are working overtime.

As time passes and input prices are free to change, the high demand will start to raise input prices. In particular, overworked employees will demand and receive raises as employers scramble to deal with the labor shortages that arise when the economy is producing at above its full-employment output level. As input prices increase, firm profits will begin to fall. And as they decline, so does the motive firms have to produce more than the full-employment output level. This process of rising input prices and falling profits continues until the rise in input prices exactly matches the initial change in output prices (in our example, they both double). When that happens, firm profits in real terms return to their original level so that firms are once again motivated to produce at exactly the full-employment output level. This adjustment process means that in the long run the economy will produce at full employment regardless of the price level (in our example, at either $P = 100$ or $P = 200$). That is why the long-run aggregate supply curve AS_{LR} is vertical above the full-employment output level. Every possible price level on the vertical axis is associated with the economy producing at the full-employment output level in the long run once input prices adjust to exactly match changes in output prices.

Focusing on the Short Run

The immediate-short-run aggregate supply curve, the short-run aggregate supply curve, and the long-run aggregate supply curve are all important. Each curve is appropriate to situations that match their respective assumptions about the flexibility of input and output prices. In the remainder of the book, we will have several different opportunities to refer to each curve. But our focus in the rest of this chapter and the several chapters that immediately follow will be on short-run aggregate supply curves, such as the AS curve shown in Figure 30.4. Indeed, unless explicitly stated otherwise, all references to "aggregate supply" are to the AS curve and to aggregate supply in the short run.

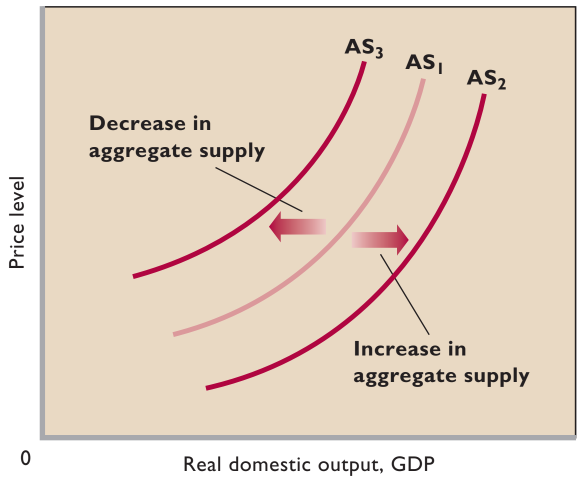
Our emphasis on the short-run aggregate supply curve AS stems from out interest in understanding the business cycle in the simplest possible way. It is a fact that real-world economies typically manifest simultaneous changes in both their price levels and their levels of real output. The upsloping short-run AS curve is the only version of aggregate supply that can handle simultaneous movements in both of these variables. By contrast, the price level is assumed fixed in the immediate-short-run version of aggregate supply illustrated in Figure 30.3 and the economy's output is always equal to the full-employment output level in the long-run version of aggregate supply shown in Figure 30.5. This renders these versions of the aggregate supply curve less useful as part of a core model for analyzing business cycles and demonstrating the short-run government policies designed to deal with them. In our current discussion, we will reserve use of the immediate short run and the long run for specific, clearly identified situations. Later in the book we will explore how the short-run and long-run AS curves are linked, and how that linkage adds several additional insights about business cycles and policy.

Changes in Aggregate Supply

LO30.4 Explain the factors that cause changes (shifts) in AS.

An existing aggregate supply curve identifies the relationship between the price level and real output, other things equal. But when one or more of these other things change, the curve itself shifts. The rightward shift of the curve

FIGURE 30.6 Changes in aggregate supply. A change in one or more of the listed determinants of aggregate supply will shift the aggregate supply curve. The rightward shift of the aggregate supply curve from AS_1 to AS_2 represents an increase in aggregate supply; the leftward shift of the curve from AS_1 to AS_3 shows a decrease in aggregate supply.



Determinants of Aggregate Supply: Factors That Shift the Aggregate Supply Curve

1. Change in input prices
   a. Domestic resource prices
   b. Prices of imported resources
2. Change in productivity
3. Change in legal-institutional environment
   a. Business taxes and subsidies
   b. Government regulations

from AS_1 to AS_2 in Figure 30.6 represents an increase in aggregate supply, indicating that firms are willing to produce and sell more real output at each price level. The leftward shift of the curve from AS_1 to AS_3 represents a decrease in aggregate supply. At each price level, firms produce less output than before.

Figure 30.6 lists the other things that cause a shift of the aggregate supply curve. Called the **determinants of aggregate supply** or *aggregate supply shifters*, they collectively position the aggregate supply curve and shift the curve when they change. Changes in these determinants raise or lower per-unit production costs *at each price level (or each level of output)*. These changes in per-unit production cost affect profits, thereby leading firms to alter the amount of output they are willing to produce *at each price level*. For example, firms may collectively offer \$9 trillion of real output at a price level of 1.0 (100 in index value), rather than \$8.8 trillion. Or they may offer \$7.5 trillion rather than \$8 trillion. The point is that when one of the determinants listed in Figure 30.6 changes, the aggregate supply curve shifts to the right or left. Changes that reduce per-unit production costs shift the aggregate supply curve to the right, as from AS_1 to AS_2; changes that increase per-unit production costs shift it to the left, as from AS_1 to AS_3. When per-unit production costs change for reasons other than changes in real output, the aggregate supply curve shifts.

The three aggregate supply determinants listed in Figure 30.6 require more discussion.

Input Prices

Input or resource prices—to be distinguished from the output prices that make up the price level—are a major ingredient of per-unit production costs and therefore a key determinant of aggregate supply. These resources can be either domestic or imported.

Domestic Resource Prices As stated earlier, wages and salaries make up about 75 percent of all business costs. Other things equal, decreases in wages reduce per-unit production costs. So when wages fall, the aggregate supply curve shifts to the right. Increases in wages shift the curve to the left. Examples:

- Labor supply increases because of substantial immigration. Wages and per-unit production costs fall, shifting the AS curve to the right.
- Labor supply decreases because a rapid increase in pension income causes many older workers to opt for early retirement. Wage rates and per-unit production costs rise, shifting the AS curve to the left.

Similarly, the aggregate supply curve shifts when the prices of land and capital inputs change. Examples:

- The price of machinery and equipment falls because of declines in the prices of steel and electronic components. Per-unit production costs decline, and the AS curve shifts to the right.
- The supply of available land resources expands through discoveries of mineral deposits, irrigation of land, or technical innovations that transform "nonresources" (say, vast desert lands) into valuable resources (productive lands). The price of land declines, per-unit production costs fall, and the AS curve shifts to the right.

Prices of Imported Resources Just as foreign demand for U.S. goods contributes to U.S. aggregate demand, resources imported from abroad (such as oil, tin, and copper) add to U.S. aggregate supply. Added supplies of resources—whether domestic or imported—typically reduce per-unit production costs. A decrease in the price of imported resources increases U.S. aggregate supply, while an increase in their price reduces U.S. aggregate supply.

A good example of the major effect that changing resource prices can have on aggregate supply is the oil price hikes of the 1970s. At that time, a group of oil-producing nations called the Organization of Petroleum Exporting Countries (OPEC) worked in concert to decrease oil production in order to raise the price of oil. The 10-fold increase in the price of oil that OPEC achieved during the 1970s drove per-unit production costs up and jolted the U.S. aggregate supply curve leftward. By contrast, a sharp decline in oil prices in the mid-1980s resulted in a rightward shift of the U.S. aggregate supply curve. In 1999 OPEC again reasserted itself, raising oil prices and therefore per-unit production costs for some U.S. producers including airlines and shipping companies like FedEx and UPS. In 2007 the price of oil shot upward, but this increase was attributed to greater demand rather than to decreases in supply caused by OPEC. But keep in mind that no matter what their cause, increases in the price of oil and other resources raise production costs and decrease aggregate supply.

Exchange-rate fluctuations are another factor that may alter the price of imported resources. Suppose that the dollar appreciates, enabling U.S. firms to obtain more foreign currency with each dollar. This means that domestic producers face a lower *dollar* price of imported resources. U.S. firms will respond by increasing their imports of foreign resources, thereby lowering their per-unit production costs at each level of output. Falling per-unit production costs will shift the U.S. aggregate supply curve to the right.

A depreciation of the dollar will have the opposite set of effects and will shift the aggregate supply curve to the left.

Productivity

The second major determinant of aggregate supply is **productivity,** which is a measure of the relationship between a nation's level of real output and the amount of resources used to produce that output. Productivity is a measure of average real output, or of real output per unit of input:

$$\text{Productivity} = \frac{\text{total output}}{\text{total inputs}}$$

An increase in productivity enables the economy to obtain more real output from its limited resources. An increase in productivity affects aggregate supply by reducing the per-unit cost of output (per-unit production cost). Suppose, for example, that real output is 10 units, that 5 units of input are needed to produce that quantity, and that the price of each input unit is $2. Then

$$\text{Productivity} = \frac{\text{total output}}{\text{total inputs}} = \frac{10}{5} = 2$$

and

$$\begin{aligned}\text{Per-unit production cost} &= \frac{\text{total input cost}}{\text{total output}}\\ &= \frac{\$2 \times 5}{10} = \$1\end{aligned}$$

Note that we obtain the total input cost by multiplying the unit input cost by the number of inputs used.

Now suppose productivity increases so that real output doubles to 20 units, while the price and quantity of the input remain constant at $2 and 5 units. Using the above equations, we see that productivity rises from 2 to 4 and that the per-unit production cost of the output falls from $1 to $0.50. The doubled productivity has reduced the per-unit production cost by half.

WORKED PROBLEMS

W30.1
Productivity and costs

By reducing the per-unit production cost, an increase in productivity shifts the aggregate supply curve to the right. The main source of productivity advance is improved production technology, often embodied within new plant and equipment that replaces old plant and equipment. Other sources of productivity increases are a better-educated and better-trained workforce, improved forms of business enterprises, and the reallocation of labor resources from lower-productivity to higher-productivity uses.

Much rarer, decreases in productivity increase per-unit production costs and therefore reduce aggregate supply (shift the curve to the left).

Legal-Institutional Environment

Changes in the legal-institutional setting in which businesses operate are the final determinant of aggregate supply. Such changes may alter the per-unit costs of output and, if so, shift the aggregate supply curve. Two changes of this type are (1) changes in taxes and subsidies and (2) changes in the extent of regulation.

Business Taxes and Subsidies Higher business taxes, such as sales, excise, and payroll taxes, increase per-unit costs and reduce short-run aggregate supply in much the same way as a wage increase does. An increase in such taxes paid by businesses will increase per-unit production costs and shift aggregate supply to the left.

Similarly, a business subsidy—a payment or tax break by government to producers—lowers production costs and increases short-run aggregate supply. For example, the federal government subsidizes firms that blend ethanol (derived from corn) with gasoline to increase the U.S. gasoline supply. This reduces the per-unit production cost of making blended gasoline. To the extent that this and other subsidies are successful, the aggregate supply curve shifts rightward.

Government Regulation It is usually costly for businesses to comply with government regulations. More regulation therefore tends to increase per-unit production costs and shift the aggregate supply curve to the left. "Supply-side" proponents of deregulation of the economy have argued forcefully that, by increasing efficiency and reducing the paperwork associated with complex regulations, deregulation will reduce per-unit costs and shift the aggregate supply curve to the right. Other economists are less certain. Deregulation that results in accounting manipulations, monopolization, and business failures is likely to shift the AS curve to the left rather than to the right.

QUICK REVIEW 30.2

- The immediate-short-run aggregate supply curve is horizontal; given fixed input and output prices, producers will supply whatever quantity of real output is demanded at the current price level.
- The short-run aggregate supply curve (or simply the "aggregate supply curve") is upsloping; given fixed resource prices, higher output prices raise firm profits and encourage them to increase their output levels.
- The long-run aggregate supply curve is vertical; given sufficient time, wages and other input prices rise or fall to match any change in the price level (that is, any change in the level of *output* prices).
- By altering per-unit production costs independent of changes in the level of output, changes in one or more of the determinants of aggregate supply (Figure 30.6) shift the aggregate supply curve.
- An increase in short-run aggregate supply is shown as a rightward shift of the aggregate supply curve; a decrease is shown as a leftward shift of the curve.

Equilibrium in the AD-AS Model

LO30.5 Discuss how AD and AS determine an economy's equilibrium price level and level of real GDP.

Of all the possible combinations of price levels and levels of real GDP, which combination will the economy gravitate toward, at least in the short run? **Figure 30.7 (Key Graph)** and its accompanying table provide the answer. Equilibrium occurs at the price level that equalizes the amounts of real output demanded and supplied. The intersection of the aggregate demand curve AD and the aggregate supply curve AS establishes the economy's **equilibrium price level** and **equilibrium real output.** So aggregate demand and aggregate supply *jointly* establish the price level and level of real GDP.

In Figure 30.7 the equilibrium price level and level of real output are 100 and $510 billion, respectively. To illustrate why, suppose the price level is 92 rather than 100. We see from the table that the lower price level will encourage businesses to produce real output of $502 billion. This is shown by point *a* on the AS curve in the graph. But, as revealed by the table and point *b* on the aggregate demand curve, buyers will want to purchase $514 billion of real output at price level 92. Competition among buyers to purchase the lesser available real output of $502 billion will eliminate the $12 billion (= $514 billion − $502 billion) shortage and pull up the price level to 100.

As the table and graph show, the rise in the price level from 92 to 100 encourages producers to increase their real output from $502 billion to $510 billion and causes buyers to scale back their purchases from $514 billion to $510 billion. When equality occurs between the amounts of real output produced and purchased, as it does at price level 100, the economy has achieved equilibrium (here, at $510 billion of real GDP).

A final note: Although the equilibrium price level happens to be 100 in our example, nothing special is implied by that. Any price level can be an equilibrium price level.

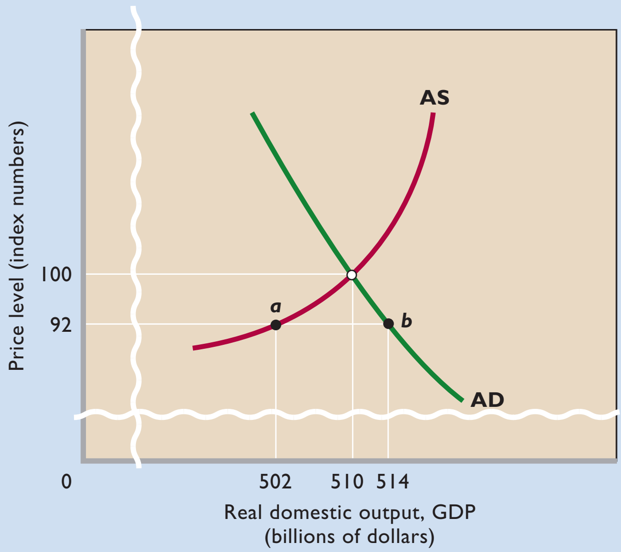
Changes in Equilibrium

LO30.6 Describe how the AD-AS model explains periods of demand-pull inflation, cost-push inflation, and recession.

Now let's apply the AD-AS model to various situations that can confront the economy. For simplicity we will use *P* and *Q* symbols, rather than actual numbers. Remember that these symbols represent, respectively, price index values and amounts of real GDP.

KEY GRAPH

FIGURE 30.7 The equilibrium price level and equilibrium real GDP. The intersection of the aggregate demand curve and the aggregate supply curve determines the economy's equilibrium price level. At the equilibrium price level of 100 (in index-value terms), the \$510 billion of real output demanded matches the \$510 billion of real output supplied. So the equilibrium GDP is \$510 billion.



Real Output Demanded (Billions)	Price Level (Index Number)	Real Output Supplied (Billions)
\$506	108	\$513
508	104	512
510	**100**	**510**
512	96	507
514	92	502

QUICK QUIZ FOR FIGURE 30.7

1. The AD curve slopes downward because:
   - **a.** per-unit production costs fall as real GDP increases.
   - **b.** the income and substitution effects are at work.
   - **c.** changes in the determinants of AD alter the amounts of real GDP demanded at each price level.
   - **d.** decreases in the price level give rise to real-balances effects, interest-rate effects, and foreign purchases effects that increase the amount of real GDP demanded.
2. The AS curve slopes upward because:
   - **a.** per-unit production costs rise as real GDP expands toward and beyond its full-employment level.
   - **b.** the income and substitution effects are at work.
   - **c.** changes in the determinants of AS alter the amounts of real GDP supplied at each price level.
   - **d.** increases in the price level give rise to real-balances effects, interest-rate effects, and foreign purchases effects that increase the amounts of real GDP supplied.
3. At price level 92:
   - **a.** a GDP surplus of \$12 billion occurs that drives the price level up to 100.
   - **b.** a GDP shortage of \$12 billion occurs that drives the price level up to 100.
   - **c.** the aggregate amount of real GDP demanded is less than the aggregate amount of real GDP supplied.
   - **d.** the economy is operating beyond its capacity to produce.
4. Suppose real output demanded rises by \$4 billion at each price level. The new equilibrium price level will be:
   - **a.** 108.
   - **b.** 104.
   - **c.** 96.
   - **d.** 92.

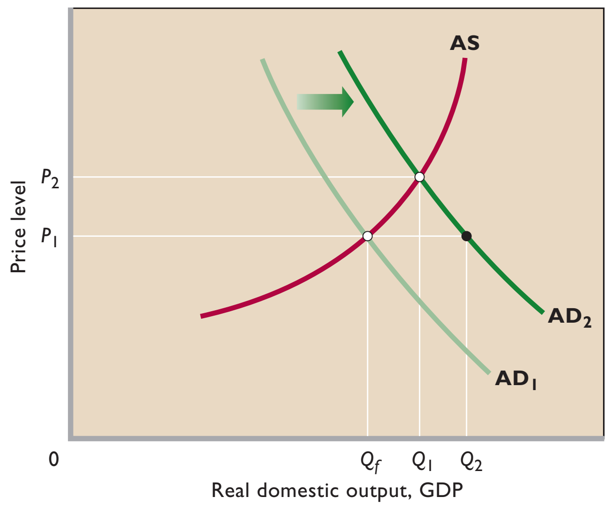
Answers: 1. d; 2. a; 3. b; 4. b

Increases in AD: Demand-Pull Inflation

Suppose the economy is operating at its full-employment output and businesses and government decide to increase their spending—actions that shift the aggregate demand curve to the right. Our list of determinants of aggregate demand (Figure 30.2) provides several reasons why this shift might occur. Perhaps firms boost their investment spending because they anticipate higher future profits from investments in new capital. Those profits are predicated on having new equipment and facilities that incorporate a number of new technologies. And perhaps government increases spending to expand national defense.

As shown by the rise in the price level from P_1 to P_2 in Figure 30.8, the increase in aggregate demand beyond the full-employment level of output causes inflation. This is

FIGURE 30.8 An increase in aggregate demand that causes demand-pull inflation. The increase of aggregate demand from AD_1 to AD_2 causes demand-pull inflation, shown as the rise in the price level from P_1 to P_2. It also causes an inflationary GDP gap of Q_1 minus Q_f. The rise of the price level reduces the size of the multiplier effect. If the price level had remained at P_1, the increase in aggregate demand from AD_1 to AD_2 would increase output from Q_f to Q_2 and the multiplier would have been at full strength. But because of the increase in the price level, real output increases only from Q_f to Q_1 and the multiplier effect is reduced.



demand-pull inflation because the price level is being pulled up by the increase in aggregate demand. Also, observe that the increase in demand expands real output from the full-employment level Q_f to Q_1. The distance between Q_1 and Q_f is a positive, or "inflationary," GDP gap. Actual GDP exceeds potential GDP.

The classic American example of demand-pull inflation occurred in the late 1960s. The escalation of the war in Vietnam resulted in a 40 percent increase in defense spending between 1965 and 1967 and another 15 percent increase in 1968. The rise in government spending, imposed on an already growing economy, shifted the economy's aggregate demand curve to the right, producing the worst inflation in two decades. Actual GDP exceeded potential GDP, thereby creating an inflationary GDP gap. Inflation jumped from 1.6 percent in 1965 to 5.7 percent by 1970.

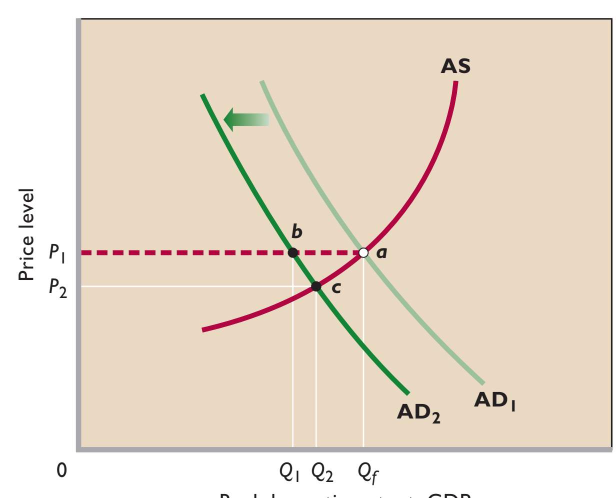
A careful examination of Figure 30.8 reveals an interesting point concerning the multiplier effect. The increase in aggregate demand from AD_1 to AD_2 increases real output only to Q_1, not to Q_2, because part of the increase in aggregate demand is absorbed as inflation as the price level rises from P_1 to P_2. Had the price level remained at P_1, the shift of aggregate demand from AD_1 to AD_2 would have increased real output to Q_2. The full-strength multiplier effect of Chapters 28 and 29 would have occurred. But in Figure 30.8 inflation reduced the increase in real output—and thus the multiplier effect—by about one-half. For any initial increase in aggregate demand, the resulting increase in real output will be smaller the greater is the increase in the price level. Price-level flexibility weakens the realized multiplier effect.

Decreases in AD: Recession and Cyclical Unemployment

Decreases in aggregate demand describe the opposite end of the business cycle: recession and cyclical unemployment (rather than above-full employment and demand-pull inflation). For example, in 2008 investment spending in the United States greatly declined because of sharply lower expected returns on investment. These lower expectations resulted from the prospects of poor future business conditions and high degrees of current unused production capacity. In Figure 30.9 we show the resulting decline in aggregate demand as a leftward shift from AD_1 to AD_2.

But we now add an important twist to the analysis: *Deflation*—a decline in the price level—is not the norm in the American economy. We discussed "sticky prices" in Chapter 24 and previously explained how fixed prices lead to horizontal immediate-short-run aggregate supply

FIGURE 30.9 A decrease in aggregate demand that causes a recession. If the price level is downwardly inflexible at P_1, a decline of aggregate demand from AD_1 to AD_2 will move the economy leftward from *a* to *b* along the horizontal broken-line segment and reduce real GDP from Q_f to Q_1. Idle production capacity, cyclical unemployment, and a recessionary GDP gap (of Q_1 minus Q_f) will result. If the price level were flexible downward, the decline in aggregate demand would move the economy depicted from *a* to *c* instead of from *a* to *b*.



curves. For reasons we will examine soon, many important prices in the U.S. economy are downwardly inflexible such that the price level is sticky downward even when aggregate demand substantially declines. Consider Figure 30.9, where the aggregate demand declines from AD_1 to AD_2. If the price level is stuck at P_1, the economy moves from *a* to *b* along the broken horizontal line rather than from *a* to *c* along the short-run aggregate supply curve AS. The outcome is a decline of real output from Q_f to Q_1, with *no* change in the price level. In this case, it is as if the aggregate supply curve in Figure 30.9 is horizontal at P_1, to the left of Q_f, as indicated by the dashed line. This decline of real output from Q_f to Q_1 constitutes a *recession*, and since fewer workers are needed to produce the lower output, *cyclical unemployment* arises. The distance between Q_1 and Q_f is a negative, or "recessionary," GDP gap—the amount by which actual output falls short of potential output.

Close inspection of Figure 30.9 also reveals that without a fall in the price level, the multiplier is at full strength. With the price level stuck at P_1, real GDP decreases by $Q_f - Q_1$, which matches the full leftward shift of the AD curve. The multiplier of Chapters 28 and 29 is at full strength when changes in aggregate demand occur along what, in effect, is a horizontal segment of the AS curve. This full-strength multiplier would also exist for an increase in aggregate demand from AD_2 to AD_1 along this broken line, since none of the increase in output would be dissipated as inflation. We will say more about that in Chapter 31.

All recent recessions in the United States have generally mimicked the "GDP gap but no deflation" scenario shown in Figure 30.9. Consider the recession of 2007–2009, which resulted from chaos in the financial markets that quickly led to significant declines in spending by businesses and households. Because of the resulting decline in aggregate demand, real GDP fell short of potential real GDP by roughly \$300 billion in 2008 and \$1 trillion in 2009. The nation's unemployment rate rose from 4.7 percent in December 2007 to 10.1 percent in October 2009. The price level fell in some months and the rate of inflation declined—meaning that *disinflation* occurred. Considering the full period, however, *deflation* did not occur.

Real output takes the brunt of declines in aggregate demand in the U.S. economy because the price level tends to be downwardly rigid in the immediate short run. There are several reasons for this downward price stickiness.

- ***Fear of price wars*** Some large firms may be concerned that if they reduce their prices, rivals not only will match their price cuts but may retaliate by making even deeper cuts. An initial price cut may touch off an unwanted *price war:* successively deeper and deeper rounds of price cuts. In such a situation, each firm eventually ends up with far less profit or higher losses than would be the case if each had simply maintained its prices. For this reason, each firm may resist making the initial price cut, choosing instead to reduce production and lay off workers.
- ***Menu costs*** Firms that think a recession will be relatively short-lived may be reluctant to cut their prices. One reason is what economists metaphorically call **menu costs,** named after their most obvious example: the cost of printing new menus when a restaurant decides to reduce its prices. But lowering prices also creates other costs. Additional costs derive from (1) estimating the magnitude and duration of

CONSIDER THIS . . .

Ratchet Effect

A *ratchet analogy* is a good way to think about the effects of changes in aggregate demand on the price level. A ratchet is a tool or mechanism such as a winch, car jack, or socket wrench that cranks a wheel forward but does not allow it to go backward. Properly set, each allows the operator to move an object (boat, car, or nut) in one direction while preventing it from moving in the opposite direction.

Product prices, wage rates, and per-unit production costs are highly flexible upward when aggregate demand increases along the aggregate supply curve. In the United States, the price level has increased in 60 of the 62 years since 1950.

But when aggregate demand decreases, product prices, wage rates, and per-unit production costs are inflexible downward. The U.S. price level has declined in only two years (1955 and 2009) since 1950, even though aggregate demand and real output have declined in a number of years.

In terms of our analogy, increases in aggregate demand ratchet the U.S. price level upward. Once in place, the higher price level remains until it is ratcheted up again. The higher price level tends to remain even with declines in aggregate demand.

the shift in demand to determine whether prices should be lowered, (2) repricing items held in inventory, (3) printing and mailing new catalogs, and (4) communicating new prices to customers, perhaps through advertising. When menu costs are present, firms may choose to avoid them by retaining current prices. That is, they may wait to see if the decline in aggregate demand is permanent.

- ***Wage contracts*** Firms rarely profit from cutting their product prices if they cannot also cut their wage rates. Wages are usually inflexible downward because large parts of the labor force work under contracts prohibiting wage cuts for the duration of the contract. (Collective bargaining agreements in major industries frequently run for 3 years. Similarly, the wages and salaries of nonunion workers are usually adjusted once a year, rather than quarterly or monthly.)
- ***Morale, effort, and productivity*** Wage inflexibility downward is reinforced by the reluctance of many employers to reduce wage rates. Some current wages may be so-called **efficiency wages**—wages that elicit maximum work effort and thus minimize labor costs per unit of output. If worker productivity (output per hour of work) remains constant, lower wages *do* reduce labor costs per unit of output. But lower wages might impair worker morale and work effort, thereby reducing productivity. Considered alone, lower productivity raises labor costs per unit of output because less output is produced. If the higher labor costs resulting from reduced productivity exceed the cost savings from the lower wage, then wage cuts will increase rather than reduce labor costs per unit of output. In such situations, firms will resist lowering wages when they are faced with a decline in aggregate demand.

ORIGIN OF THE IDEA

O30.2

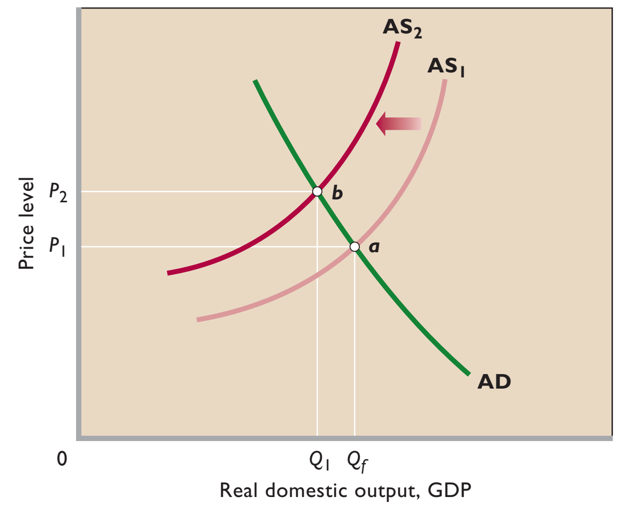
Efficiency wage

- ***Minimum wage*** The minimum wage imposes a legal floor under the wages of the least-skilled workers. Firms paying those wages cannot reduce that wage rate when aggregate demand declines.

Decreases in AS: Cost-Push Inflation

Suppose that a major terrorist attack on oil facilities severely disrupts world oil supplies and drives up oil prices by, say, 300 percent. Higher energy prices would spread through the economy, driving up production and distribution costs on a wide variety of goods. The U.S. aggregate supply curve would shift to the left, say, from AS_1 to AS_2 in Figure 30.10. The resulting increase in the price level would be *cost-push inflation*.

FIGURE 30.10 A decrease in aggregate supply that causes cost-push inflation. A leftward shift of aggregate supply from AS_1 to AS_2 raises the price level from P_1 to P_2 and produces cost-push inflation. Real output declines and a recessionary GDP gap (of Q_1 minus Q_f) occurs.

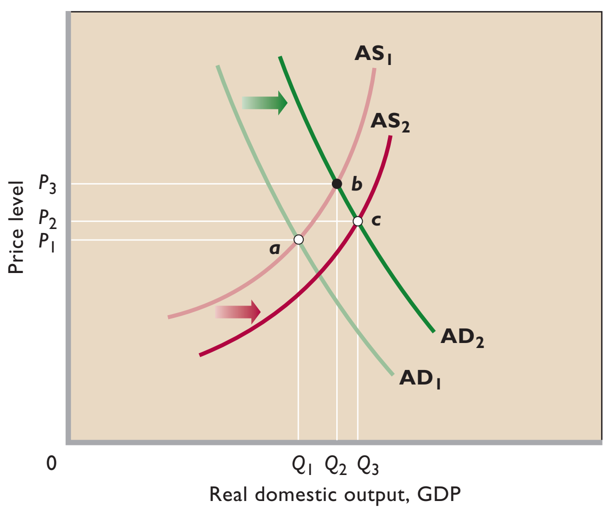


The effects of a leftward shift in aggregate supply are doubly bad. When aggregate supply shifts from AS_1 to AS_2, the economy moves from *a* to *b*. The price level rises from P_1 to P_2 and real output declines from Q_f to Q_1. Along with the cost-push inflation, a recession (and negative GDP gap) occurs. That is exactly what happened in the United States in the mid-1970s when the price of oil rocketed upward. Then, oil expenditures were about 10 percent of U.S. GDP, compared to only 3 percent today. So the U.S. economy is now less vulnerable to cost-push inflation arising from such "aggregate supply shocks." That said, it is not *immune* from such shocks.

Increases in AS: Full Employment with Price-Level Stability

Between 1996 and 2000, the United States experienced a combination of full employment, strong economic growth, and very low inflation. Specifically, the unemployment rate fell to 4 percent and real GDP grew nearly 4 percent annually, *without igniting inflation*. At first thought, this "macroeconomic bliss" seems to be incompatible with the

FIGURE 30.11 Growth, full employment, and relative price stability. Normally, an increase in aggregate demand from AD_1 to AD_2 would move the economy from *a* to *b* along AS_1. Real output would expand to Q_2, and inflation would result (P_1 to P_3). But in the late 1990s, significant increases in productivity shifted the aggregate supply curve, as from AS_1 to AS_2. The economy moved from *a* to *c* rather than from *a* to *b*. It experienced strong economic growth (Q_1 to Q_3), full employment, and only very mild inflation (P_1 to P_2) before receding in March 2001.



AD-AS model. The aggregate supply curve suggests that increases in aggregate demand that are sufficient for over-full employment will raise the price level (see Figure 30.8). Higher inflation, so it would seem, is the inevitable price paid for expanding output beyond the full-employment level.

But inflation remained very mild in the late 1990s. Figure 30.11 helps explain why. Let's first suppose that aggregate demand increased from AD_1 to AD_2 along aggregate supply curve AS_1. Taken alone, that increase in aggregate demand would move the economy from *a* to *b*. Real output would rise from full-employment output Q_1 to beyond-full-employment output Q_2. The economy would experience inflation, as shown by the increase in the price level from P_1 to P_3. Such inflation had occurred at the end of previous vigorous expansions of aggregate demand, including the expansion of the late 1980s.

Between 1990 and 2000, however, larger-than-usual increases in productivity occurred because of a burst of new technology relating to computers, the Internet, inventory management systems, electronic commerce, and so on. We represent this higher-than-usual productivity growth as the rightward shift from AS_1 to AS_2 in Figure 30.11. The relevant aggregate demand and aggregate supply curves thus became AD_2 and AS_2, not AD_2 and AS_1. Instead of moving from *a* to *b*, the economy moved from *a* to *c*. Real output increased from Q_1 to Q_3, and the price level rose only modestly (from P_1 to P_2). The shift of the aggregate supply curve from AS_1 to AS_2 accommodated the rapid increase in aggregate demand and kept inflation mild. This remarkable combination of rapid productivity growth, rapid real GDP growth, full employment, and relative price-level stability led some observers to proclaim that the United States was experiencing a "new era" or a New Economy.

But in 2001 the New Economy came face-to-face with the old economic principles. Aggregate demand declined because of a substantial fall in investment spending, and in March 2001 the economy experienced a recession. The terrorist attacks of September 11, 2001, further dampened private spending and prolonged the recession throughout 2001. The unemployment rate rose from 4.2 percent in January 2001 to 6 percent in December 2002.

The economy rebounded between 2002 and 2007, eventually reachieving its earlier strong economic growth, low inflation, and low unemployment. Some economists began to refer to the period after 1982 as "The Great Moderation" because recessions were father apart and relatively mild. They drew the implication that businesses and government had smoothed out the business cycle. Wrong! The severity of the recession of 2007–2009 was a huge surprise to most economists. And so was the weakness of the subsequent recovery, as discussed in this chapter's Last Word.

QUICK REVIEW 30.3

- The equilibrium price level and amount of real output are determined at the intersection of the aggregate demand curve and the aggregate supply curve.
- Increases in aggregate demand beyond the full-employment level of real GDP cause demand-pull inflation.
- Decreases in aggregate demand cause recessions and cyclical unemployment, partly because the price level and wages tend to be inflexible in a downward direction.
- Decreases in aggregate supply cause cost-push inflation.
- Full employment, high economic growth, and price stability are compatible with one another if productivity-driven increases in aggregate supply are sufficient to balance growing aggregate demand.

LAST WORD

Stimulus and the Great Recession

Aggregate Demand Stimulus Helped to Prevent the 2007–2009 Downturn from Becoming Another Great Depression. But Why Was the Stimulus-Fueled Recovery Substantially Weaker Than Expected?

In retrospect, it is clear that the U.S. economy was in a precarious position in 2006. Trillions of dollars had been borrowed to buy housing on the expectation that home prices would keep on rising. That expectation made borrowing seem like a "no brainer" as a potential buyer could anticipate that if she borrowed $200,000 to buy a house in one year, she would be able to sell it the next year for, say, $215,000. Selling at a higher price would allow her to pay off the $200,000 loan and keep the rest as pure profit.

Unfortunately, home prices started to fall in 2006. When they did, many people who had borrowed to buy houses found themselves unable to pay off their loans. That in turn meant that many banks found themselves holding loans that would never be paid back. Soon, many banks teetered on bankruptcy, the financial markets began to freeze up, and it became clear by late 2007 that the overall economy would probably enter a recession as the result of the housing collapse.

When it was widely recognized in late 2008 that the downturn was going to be unusually severe, public officials took extraordinarily strong steps to stimulate aggregate demand. In terms of monetary policy, the Federal Reserve lowered short-term interest rates to nearly zero in order to shift AD to the right by stimulating investment and consumption. In terms of fiscal policy, the federal government began the country's largest peacetime program of deficit-funded spending increases. Those spending increases also shifted AD to the right by increasing the total amount of government expenditures.

Those actions were widely credited with preventing a much worse downturn. Real GDP did fall by 4.7 percent and

SUMMARY

LO30.1 Define aggregate demand (AD) and explain how its downward slope is the result of the real-balances effect, the interest-rate effect, and the foreign purchases effect.

The aggregate demand–aggregate supply model (AD-AS model) is a flexible-price model that enables analysis of simultaneous changes of real GDP and the price level.

The aggregate demand curve shows the level of real output that the economy demands at each price level.

The aggregate demand curve is downsloping because of the real-balances effect, the interest-rate effect, and the foreign purchases effect. The real-balances effect indicates that inflation reduces the real value or purchasing power of fixed-value financial assets held by households, causing cutbacks in consumer spending. The interest-rate effect means that, with a specific supply of money, a higher price level increases the demand for money, thereby raising the interest rate and reducing investment purchases. The foreign purchases effect suggests that an increase in one country's price level relative to the price levels in other countries reduces the net export component of that nation's aggregate demand.

LO30.2 Explain the factors that cause changes (shifts) in AD.

The determinants of aggregate demand consist of spending by domestic consumers, by businesses, by government, and by foreign buyers. Changes in the factors listed in Figure 30.2 alter the spending by these groups and shift the aggregate demand curve. The extent of the shift is determined by the size of the initial change in spending and the strength of the economy's multiplier.

the unemployment rate did rise from 4.6 to 10.1 percent. But those negative changes were much less severe than what had happened during the Great Depression of the 1930s, when real GDP fell by nearly 27 percent and the unemployment rate rose to nearly 25 percent.

As time passed, however, it became clear that the stimulus was having less of an effect than many economists had anticipated. White House economists, for instance, had predicted that the stimulus begun in 2009 would reduce the unemployment rate to 5.2 percent by 2012. But three years later the unemployment rate was still at 7.8 percent despite the Federal Reserve continuing to keep interest rates extremely low and despite the federal government continuing to run massive deficits to fund huge amounts of government expenditures.

GDP growth was also disappointing. Real GDP expanded by only 2.4 percent in 2010, 1.8 percent in 2011, and 1.6 percent in 2012. By contrast, the period after the early 1980s recession had seen annual growth rates as high as 7.2 percent per year.

One explanation for the disappointing unemployment and GDP numbers was that it was hard for the stimulus to be very effective given the high debt levels that were built up during the bubble years. The lower interest rates engineered by the Federal Reserve, for instance, were probably not much of an inducement for consumers to increase their borrowing when so many of them were already heavily in debt.

A related problem was that savings rates had risen. When the government attempted to use deficit spending and fiscal policy to stimulate the economy, policy makers were hoping that each dollar of government spending would induce many dollars of consumer spending. But debt-strapped consumers were devoting large parts of their income to making interest payments on debt or paying off loans. So, when stimulus dollars came their way, they often short-circuited the spending process by saving a lot rather than spending a lot.

Another issue was that the stimulus was diffuse while the sectors of the economy in greatest need of stimulus were focused. In particular, the government's stimulus efforts shifted *aggregate* demand to the right. But not all sectors had been hit equally hard by the recession. Thus, when AD shifted right, a lot of the effect was felt in business sectors that hadn't been hurt that badly during the recession. Meanwhile, many sectors that had been hit hard only received a small portion of the total amount of stimulus that they would have needed to see a full recovery.

A related problem is that in some sectors of the economy, the government's stimulus may have resulted mostly in price increases rather than output gains. That is because the supply curves for many industries are steep. Consider dentists and jewelers. It takes many years to train competent dentists or skilled jewelers. So even if the demand for their services shifts right, there is a nearly fixed supply of dental services and jewelry services in the short run—meaning that any increase in demand will mostly cause higher prices rather than higher output. So when the government shifted *aggregate* demand to the right, certain sectors probably saw mostly price increases rather than output gains.

LO30.3 Define aggregate supply (AS) and explain how it differs in the immediate short run, the short run, and the long run.

The aggregate supply curve shows the levels of real output that businesses will produce at various possible price levels. The slope of the aggregate supply curve depends upon the flexibility of input and output prices. Since these vary over time, aggregate supply curves are categorized into three time horizons, each having different underlying assumptions about the flexibility of input and output prices.

The *immediate-short-run aggregate supply curve* assumes that both input prices and output prices are fixed. With output prices fixed, the aggregate supply curve is a horizontal line at the current price level. The *short-run aggregate supply curve* assumes nominal wages and other input prices remain fixed while output prices vary. The aggregate supply curve is generally upsloping because per-unit production costs, and hence the prices that firms must receive, rise as real output expands. The aggregate supply curve is relatively steep to the right of the full-employment output level and relatively flat to the left of it. The *long-run aggregate supply curve* assumes that nominal wages and other input prices fully match any change in the price level. The curve is vertical at the full-employment output level.

Because the short-run aggregate supply curve is the only version of aggregate supply that can handle simultaneous changes in the price level and real output, it serves well as the core aggregate supply curve for analyzing the business cycle and economic policy. Unless stated otherwise, all references to "aggregate supply" refer to short-run aggregate supply and the short-run aggregate supply curve.

LO30.4 Explain the factors that cause changes (shifts) in AS.

Figure 30.6 lists the determinants of aggregate supply: input prices, productivity, and the legal-institutional environment. A change in any one of these factors will change per-unit production costs at each level of output and therefore will shift the aggregate supply curve.

LO30.5 Discuss how AD and AS determine an economy's equilibrium price level and level of real GDP.

The intersection of the aggregate demand and aggregate supply curves determines an economy's equilibrium price level and real GDP. At the intersection, the quantity of real GDP demanded equals the quantity of real GDP supplied.

LO30.6 Describe how the AD-AS model explains periods of demand-pull inflation, cost-push inflation, and recession.

Increases in aggregate demand to the right of the full-employment output cause inflation and positive GDP gaps (actual GDP exceeds potential GDP). An upsloping aggregate supply curve weakens the multiplier effect of an increase in aggregate demand because a portion of the increase in aggregate demand is dissipated in inflation.

Shifts of the aggregate demand curve to the left of the full-employment output cause recession, negative GDP gaps, and cyclical unemployment. The price level may not fall during recessions because of downwardly inflexible prices and wages. This inflexibility results from fear of price wars, menu costs, wage contracts, efficiency wages, and minimum wages. When the price level is fixed, changes in aggregate demand produce full-strength multiplier effects.

Leftward shifts of the aggregate supply curve reflect increases in per-unit production costs and cause cost-push inflation, with accompanying negative GDP gaps.

Rightward shifts of the aggregate supply curve, caused by large improvements in productivity, help explain the simultaneous achievement of full employment, economic growth, and price stability that occurred in the United States between 1996 and 2000. The recession of 2001, however, ended the expansionary phase of the business cycle. Expansion resumed in the 2002–2007 period, before giving way to the severe recession of 2007–2009.

TERMS AND CONCEPTS

aggregate demand–aggregate supply (AD-AS) model
aggregate demand
real-balances effect
interest-rate effect
foreign purchases effect
determinants of aggregate demand
aggregate supply
immediate-short-run aggregate supply curve
short-run aggregate supply curve
long-run aggregate supply curve
determinants of aggregate supply
productivity
equilibrium price level
equilibrium real output
menu costs
efficiency wages

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Why is the aggregate demand curve downsloping? Specify how your explanation differs from the explanation for the downsloping demand curve for a single product. What role does the multiplier play in shifts of the aggregate demand curve? **LO30.1**
2. Distinguish between "real-balances effect" and "wealth effect," as the terms are used in this chapter. How does each relate to the aggregate demand curve? **LO30.1**
3. What assumptions cause the immediate-short-run aggregate supply curve to be horizontal? Why is the long-run aggregate supply curve vertical? Explain the shape of the short-run aggregate supply curve. Why is the short-run aggregate supply curve relatively flat to the left of the full-employment output and relatively steep to the right? **LO30.3**
4. Explain how an upsloping aggregate supply curve weakens the realized multiplier effect from an initial change in investment spending. **LO30.6**
5. Why does a reduction in aggregate demand in the actual economy reduce real output, rather than the price level? Why might a full-strength multiplier apply to a decrease in aggregate demand? **LO30.6**
6. Explain: "Unemployment can be caused by a decrease of aggregate demand or a decrease of aggregate supply." In each case, specify the price-level outcomes. **LO30.6**
7. Use shifts of the AD and AS curves to explain (*a*) the U.S. experience of strong economic growth, full employment, and price stability in the late 1990s and early 2000s and (*b*) how a strong negative wealth effect from, say, a precipitous

drop in house prices could cause a recession even though productivity is surging. LO30.6

8. In early 2001 investment spending sharply declined in the United States. In the two months following the September 11, 2001, attacks on the United States, consumption also declined. Use AD-AS analysis to show the two impacts on real GDP. LO30.6

9. **LAST WORD** What were the monetary and fiscal policy responses to the Great Recession? What were some of the reasons suggested for why those policy responses didn't seem to have as large an effect as anticipated on unemployment and GDP growth?

REVIEW QUESTIONS

1. Which of the following help to explain why the aggregate demand curve slopes downward? LO30.1
   - **a.** When the domestic price level rises, our goods and services become more expensive to foreigners.
   - **b.** When government spending rises, the price level falls.
   - **c.** There is an inverse relationship between consumer expectations and personal taxes.
   - **d.** When the price level rises, the real value of financial assets (like stocks, bonds, and savings account balances) declines.
2. Which of the following will shift the aggregate demand curve to the left? LO30.2
   - **a.** The government reduces personal income taxes.
   - **b.** Interest rates rise.
   - **c.** The government raises corporate profit taxes.
   - **d.** There is an economic boom overseas that raises the incomes of foreign households.
3. Label each of the following descriptions as being either an immediate-short-run aggregate supply curve, a short-run aggregate supply curve, or a long-run aggregate supply curve. LO30.3
   - **a.** A vertical line.
   - **b.** The price level is fixed.
   - **c.** Output prices are flexible, but input prices are fixed.
   - **d.** A horizontal line.
   - **e.** An upsloping curve.
   - **f.** Output is fixed.
4. Which of the following will shift the aggregate supply curve to the right? LO30.4
   - **a.** A new networking technology increases productivity all over the economy.
   - **b.** The price of oil rises substantially.
   - **c.** Business taxes fall.
   - **d.** The government passes a law doubling all manufacturing wages.
5. At the current price level, producers supply $375 billion of final goods and services while consumers purchase $355 billion of final goods and services. The price level is: LO30.5
   - **a.** Above equilibrium.
   - **b.** At equilibrium.
   - **c.** Below equilibrium.
   - **d.** More information is needed.
6. What effects would each of the following have on aggregate demand or aggregate supply, other things equal? In each case, use a diagram to show the expected effects on the equilibrium price level and the level of real output, assuming that the price level is flexible both upward and downward. LO30.5
   - **a.** A widespread fear by consumers of an impending economic depression.
   - **b.** A new national tax on producers based on the value added between the costs of the inputs and the revenue received from their output.
   - **c.** A reduction in interest rates at each price level.
   - **d.** A major increase in spending for health care by the federal government.
   - **e.** The general expectation of coming rapid inflation.
   - **f.** The complete disintegration of OPEC, causing oil prices to fall by one-half.
   - **g.** A 10 percent across-the-board reduction in personal income tax rates.
   - **h.** A sizable increase in labor productivity (with no change in nominal wages).
   - **i.** A 12 percent increase in nominal wages (with no change in productivity).
   - **j.** An increase in exports that exceeds an increase in imports (not due to tariffs).
7. True or False: Decreases in AD normally lead to decreases in both output and the price level. LO30.6
8. Assume that (*a*) the price level is flexible upward but not downward and (*b*) the economy is currently operating at its full-employment output. Other things equal, how will each of the following affect the equilibrium price level and equilibrium level of real output in the short run? LO30.6
   - **a.** An increase in aggregate demand.
   - **b.** A decrease in aggregate supply, with no change in aggregate demand.
   - **c.** Equal increases in aggregate demand and aggregate supply.
   - **d.** A decrease in aggregate demand.
   - **e.** An increase in aggregate demand that exceeds an increase in aggregate supply.
9. True or False: If the price of oil suddenly increases by a large amount, AS will shift left, but the price level will not rise thanks to price inflexibility. LO30.6

PROBLEMS

1. Suppose that consumer spending initially rises by $5 billion for every 1 percent rise in household wealth and that investment spending initially rises by $20 billion for every 1 percentage point fall in the real interest rate. Also assume that the economy's multiplier is 4. If household wealth falls by 5 percent because of declining house values, and the real interest rate falls by 2 percentage points, in what direction and by how much will the aggregate demand curve initially shift at each price level? In what direction and by how much will it eventually shift? **LO30.2**
2. Answer the following questions on the basis of the following three sets of data for the country of North Vaudeville: **LO30.4**

(A)		(B)		(C)	
Price Level	Real GDP	Price Level	Real GDP	Price Level	Real GDP
110	275	100	200	110	225
100	250	100	225	100	225
95	225	100	250	95	225
90	200	100	275	90	225

   a. Which set of data illustrates aggregate supply in the immediate short run in North Vaudeville? The short run? The long run?
   b. Assuming no change in hours of work, if real output per hour of work increases by 10 percent, what will be the new levels of real GDP in the right column of A? Do the new data reflect an increase in aggregate supply or do they indicate a decrease in aggregate supply?
3. Suppose that the aggregate demand and aggregate supply schedules for a hypothetical economy are as shown in the following table. **LO30.5**

Amount of Real GDP Demanded, Billions	Price Level (Price Index)	Amount of Real GDP Supplied, Billions
$100	300	$450
200	250	400
300	200	300
400	150	200
500	100	100

   a. Use the data above to graph the aggregate demand and aggregate supply curves. What are the equilibrium price level and the equilibrium level of real output in this hypothetical economy? Is the equilibrium real output also necessarily the full-employment real output?
   b. If the price level in this economy is 150, will quantity demanded equal, exceed, or fall short of quantity supplied? By what amount? If the price level is 250, will quantity demanded equal, exceed, or fall short of quantity supplied? By what amount?
   c. Suppose that buyers desire to purchase $200 billion of extra real output at each price level. Sketch in the new aggregate demand curve as AD_1. What are the new equilibrium price level and level of real output?
4. Suppose that the table presented below shows an economy's relationship between real output and the inputs needed to produce that output: **LO30.4**

Input Quantity	Real GDP
150.0	$400
112.5	300
75.0	200

   a. What is productivity in this economy?
   b. What is the per-unit cost of production if the price of each input unit is $2?
   c. Assume that the input price increases from $2 to $3 with no accompanying change in productivity. What is the new per-unit cost of production? In what direction would the $1 increase in input price push the economy's aggregate supply curve? What effect would this shift of aggregate supply have on the price level and the level of real output?
   d. Suppose that the increase in input price does not occur but, instead, that productivity increases by 100 percent. What would be the new per-unit cost of production? What effect would this change in per-unit production cost have on the economy's aggregate supply curve? What effect would this shift of aggregate supply have on the price level and the level of real output?
5. Refer to the data in the table that accompanies problem 2. Suppose that the present equilibrium price level and level of real GDP are 100 and $225, and that data set B represents the relevant aggregate supply schedule for the economy. **LO30.6**
   a. What must be the current amount of real output demanded at the 100 price level?
   b. If the amount of output demanded declined by $25 at the 100 price levels shown in B, what would be the new equilibrium real GDP? In business cycle terminology, what would economists call this change in real GDP?

The Relationship of the Aggregate Demand Curve to the Aggregate Expenditures Model*

LO30.7 Identify how the aggregate demand curve relates to the aggregate expenditures model.

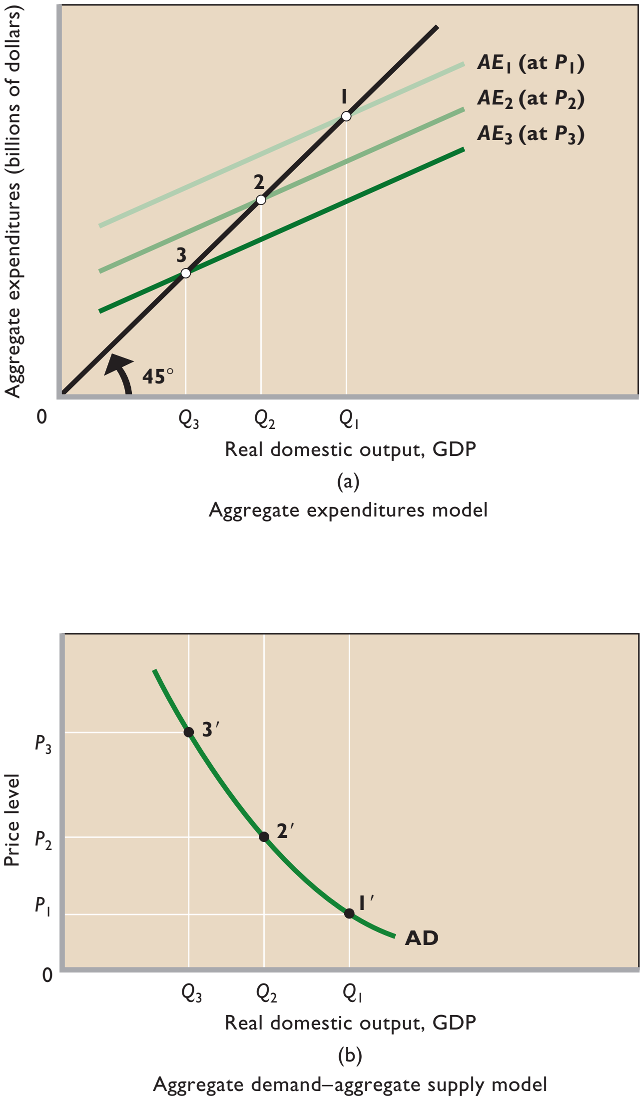
The aggregate demand curve of this chapter and the aggregate expenditures model of Chapter 29 are intricately related.

Derivation of the Aggregate Demand Curve from the Aggregate Expenditures Model

We can directly connect the downsloping aggregate demand curve to the aggregate expenditures model by relating various possible price levels to corresponding equilibrium GDPs. In Figure 1 we have stacked the aggregate expenditures model (Figure 1a) and the aggregate demand curve (Figure 1b) vertically. This is possible because the horizontal axes of both models measure real GDP. Now let's derive the AD curve in three distinct steps. (Throughout this discussion, keep in mind that price level P_1 is lower than price level P_2, which is lower than price level P_3.)

- First suppose that the economy's price level is P_1 and its aggregate expenditures schedule is AE_1, the top schedule in Figure 1a. The equilibrium GDP is then Q_1 at point 1. So in Figure 1b we can plot the equilibrium real output Q_1 and the corresponding price level P_1. This gives us point 1′ in Figure 1b.
- Now assume the price level rises from P_1 to P_2. Other things equal, this higher price level will (1) decrease the value of real balances (wealth), decreasing consumption expenditures; (2) increase the interest rate, reducing investment and interest-sensitive consumption expenditures; and (3) increase imports and decrease exports, reducing net export

FIGURE 1 Deriving the aggregate demand curve from the aggregate expenditures model. (a) Rising price levels from P_1 to P_2 to P_3 shift the aggregate expenditures curve downward from AE_1 to AE_2 to AE_3 and reduce real GDP from Q_1 to Q_2 to Q_3. (b) The aggregate demand curve AD is derived by plotting the successively lower real GDPs from the upper graph against the P_1, P_2, and P_3 price levels.



*This appendix presumes knowledge of the aggregate expenditures model discussed in Chapter 29 and should be skipped if Chapter 29 was not assigned.

expenditures. The aggregate expenditures schedule will fall from AE_1 to, say, AE_2 in Figure 1a, giving us equilibrium Q_2 at point 2. In Figure 1b we plot this new price-level–real-output combination, P_2 and Q_2, as point 2′.

- Finally, suppose the price level rises from P_2 to P_3. The value of real balances falls, the interest rate rises, exports fall, and imports rise. Consequently, the consumption, investment, and net export schedules fall, shifting the aggregate expenditures schedule downward from AE_2 to AE_3, which gives us equilibrium Q_3 at point 3. In Figure 1b, this enables us to locate point 3′, where the price level is P_3 and real output is Q_3.

In summary, increases in the economy's price level will successively shift its aggregate expenditures schedule downward and will reduce real GDP. The resulting price-level–real-GDP combinations will yield various points such as 1′, 2′, and 3′ in Figure 1b. Together, such points locate the downsloping aggregate demand curve for the economy.

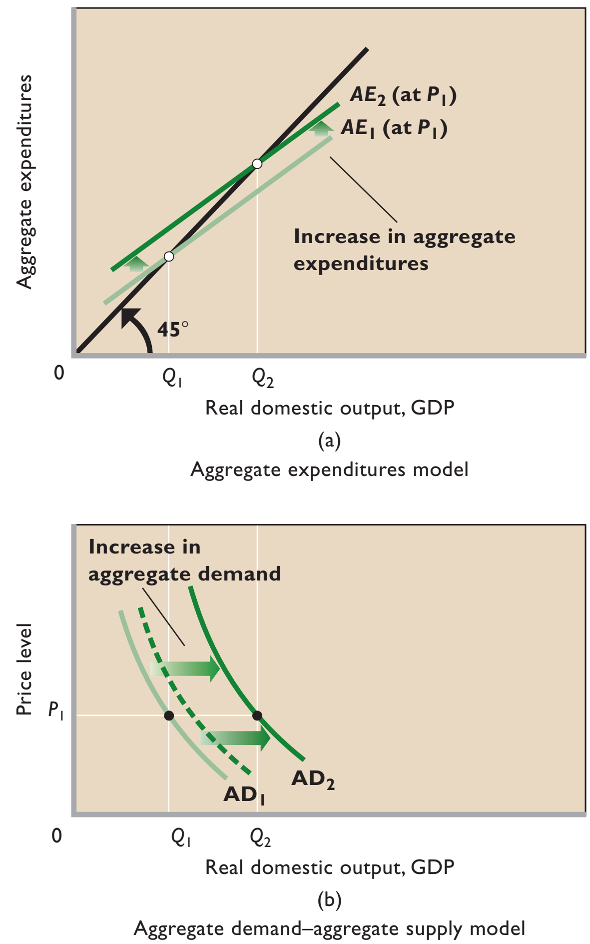
Aggregate Demand Shifts and the Aggregate Expenditures Model

The determinants of aggregate demand listed in Figure 30.2 are the components of the aggregate expenditures model discussed in Chapter 29. When one of the determinants of aggregate demand changes, the aggregate expenditures schedule shifts upward or downward. We can easily link such shifts of the aggregate expenditures schedule to shifts of the aggregate demand curve.

Let's suppose that the price level is constant. In Figure 2 we begin with the aggregate expenditures schedule at AE_1 in the top diagram, yielding equilibrium real output Q_1. Assume now that investment increases in response to more optimistic business expectations, so the aggregate expenditures schedule rises from AE_1 to AE_2. (The notation "at P_1" reminds us that the price level is assumed constant.) The result will be a multiplied increase in equilibrium real output from Q_1 to Q_2.

In Figure 2b the increase in investment spending is reflected in the horizontal distance between AD_1 and the broken curve to its right. The immediate effect of the increase in investment is an increase in aggregate demand by the exact amount of the new spending. But then the multiplier process magnifies the initial increase in investment into successive rounds of consumption spending

FIGURE 2 Shifts of the aggregate expenditures schedule and of the aggregate demand curve. (a) A change in some determinant of consumption, investment, or net exports (other than the price level) shifts the aggregate expenditures schedule upward from AE_1 to AE_2. The multiplier increases real output from Q_1 to Q_2. (b) The counterpart of this change is an initial rightward shift of the aggregate demand curve by the amount of initial new spending (from AD_1 to the broken curve). This leads to a multiplied rightward shift of the curve to AD_2, which is just sufficient to show the same increase of real output as that in the aggregate expenditures model.



and an ultimate multiplied increase in aggregate demand from AD_1 to AD_2. Equilibrium real output rises from Q_1 to Q_2, the same multiplied increase in real GDP as that in the top graph. The initial increase in investment in the top graph has shifted the AD curve in the lower graph by a horizontal distance equal to the change in investment times the multiplier. This particular change in real GDP is still associated with the constant price level P_1. To generalize,

$$\text{Shift of AD curve} = \text{initial change in spending} \times \text{multiplier}$$

APPENDIX SUMMARY

LO30.7 Identify how the aggregate demand curve relates to the aggregate expenditures model.

A change in the price level alters the location of the aggregate expenditures schedule through the real-balances, interest-rate, and foreign purchases effects. The aggregate demand curve is derived from the aggregate expenditures model by allowing the price level to change and observing the effect on the aggregate expenditures schedule and thus on equilibrium GDP.

With the price level held constant, increases in consumption, investment, government, and net export expenditures shift the aggregate expenditures schedule upward and the aggregate demand curve to the right. Decreases in these spending components produce the opposite effects.

The following and additional problems can be found in connect ECONOMICS

APPENDIX DISCUSSION QUESTIONS

1. Explain carefully: "A change in the price level shifts the aggregate expenditures curve but not the aggregate demand curve." **LO30.7**
2. Suppose that the price level is constant and that investment decreases sharply. How would you show this decrease in the aggregate expenditures model? What would be the outcome for real GDP? How would you show this fall in investment in the aggregate demand–aggregate supply model, assuming the economy is operating in what, in effect, is a horizontal section of the aggregate supply curve? **LO30.7**

APPENDIX REVIEW QUESTIONS

1. True or False: A higher price level increases aggregate expenditures. **LO30.7**
2. If the government decreases expenditures, the *AE* curve will shift ____________ and the AD curve will shift ____________. **LO30.7**
   a. Down; left.
   b. Down; right.
   c. Up; left.
   d. Up; right.

APPENDIX PROBLEMS

1. Refer to Figures 1a and 1b in the Appendix. Assume that Q_1 is 300, Q_2 is 200, Q_3 is 100, P_3 is 120, P_2 is 100, and P_1 is 80. If the price level increases from P_1 to P_3 in graph 1b, in what direction and by how much will real GDP change? If the slopes of the AE lines in Figure 1a are 0.8 and equal to the MPC, in what direction will the aggregate expenditures schedule in Figure 1a need to shift to produce the previously determined change in real GDP? What is the size of the multiplier in this example? **LO30.7**
2. Refer to Figure 2 in the Appendix and assume that Q_1 is \$400 and Q_2 is \$500, the price level is stuck at P_1, and the slopes of the AE lines in Figure 2a are 0.75 and equal to the MPC. In what direction and by how much does the aggregate expenditures schedule in Figure 2a need to shift to move the aggregate demand curve in Figure 2b from AD_1 to AD_2? What is the multiplier in this example? Given the multiplier, what must be the distance between AD_1 and the broken line to its right at P_1? **LO30.7**

CHAPTER **31**

Fiscal Policy, Deficits, and Debt

Learning Objectives

LO31.1 Identify and explain the purposes, tools, and limitations of fiscal policy.

LO31.2 Explain the role of built-in stabilizers in moderating business cycles.

LO31.3 Describe how the cyclically adjusted budget reveals the status of U.S. fiscal policy.

LO31.4 Summarize recent U.S. fiscal policy and the projections for U.S. fiscal policy over the next few years.

LO31.5 Discuss the problems that governments may encounter in enacting and applying fiscal policy.

LO31.6 Discuss the size, composition, and consequences of the U.S. public debt.

In the previous chapter we saw that an excessive increase in aggregate demand can cause demand-pull inflation and that a significant decline in aggregate demand can cause recession and cyclical unemployment. For these reasons, the federal government sometimes uses budgetary actions to try to "stimulate the economy" or "rein in inflation." Such countercyclical **fiscal policy** consists of deliberate changes in government spending and tax collections designed to achieve full employment, control inflation, and encourage economic growth. (The adjective "fiscal" simply means "financial.")

ORIGIN OF THE IDEA

O31.1

Fiscal policy

We begin this chapter by examining the logic behind fiscal policy, its current status, and its limitations. Then we examine a closely related topic: the U.S. public debt.

Our discussion of fiscal policy and public debt is very timely. In 2009, Congress and the Obama administration began a $787 billion stimulus program designed to help lift the U.S. economy out of deep recession. This fiscal policy contributed to a $1.4 trillion federal budget deficit in 2009, which increased the size of the U.S. public debt to $11.9 trillion. Large deficits continued in subsequent years, so that the U.S. public debt passed $17.0 trillion in 2013.

Fiscal Policy and the AD-AS Model

LO31.1 Identify and explain the purposes, tools, and limitations of fiscal policy.

The fiscal policy just defined is *discretionary* (or "active"). It is often initiated on the advice of the president's **Council of Economic Advisers (CEA),** a group of three economists appointed by the president to provide expertise and assistance on economic matters. Discretionary changes in government spending and taxes are *at the option* of the federal government. They do not occur automatically. Changes that occur without congressional action are *nondiscretionary* (or "passive" or "automatic"), and we will examine them later in this chapter.

Expansionary Fiscal Policy

When recession occurs, an **expansionary fiscal policy** may be in order. This policy consists of government spending increases, tax reductions, or both, designed to increase aggregate demand and therefore raise real GDP. Consider Figure 31.1, where we suppose that a sharp decline in investment spending has shifted the economy's aggregate demand curve to the left from AD_1 to AD_2. (Disregard the arrows and dashed downsloping line for now.) The cause of the recession may be that profit expectations on investment projects have dimmed, curtailing investment spending and reducing aggregate demand.

Suppose the economy's potential or full-employment output is $510 billion in Figure 31.1. If the price level is inflexible downward at P_1, the broken horizontal line becomes relevant to the analysis. The aggregate demand curve moves leftward and reduces real GDP from $510 billion to $490 billion. A negative GDP gap of $20 billion (= $490 billion − $510 billion) arises. An increase in unemployment accompanies this negative GDP gap because fewer workers are needed to produce the reduced output. In short, the economy depicted is suffering both recession and cyclical unemployment.

What fiscal policy should the federal government adopt to try to stimulate the economy? It has three main options: (1) increase government spending, (2) reduce taxes, or (3) use some combination of the two. If the federal budget is balanced at the outset, expansionary fiscal policy will create a government **budget deficit**—government spending in excess of tax revenues.

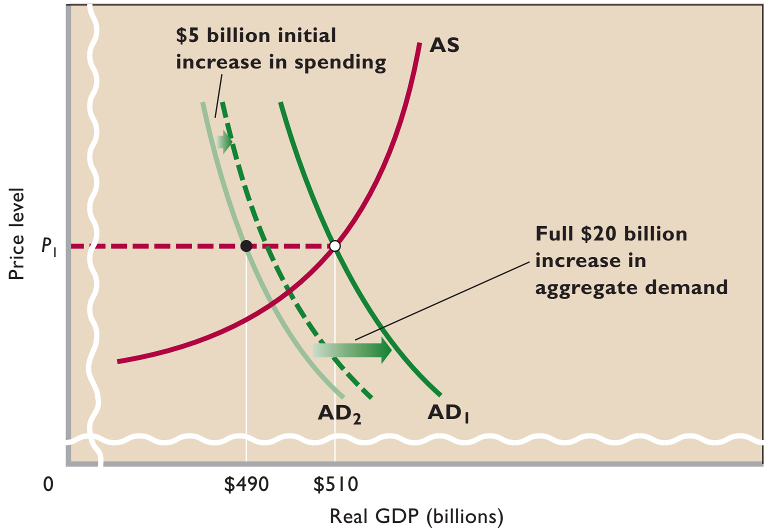


FIGURE 31.1 Expansionary fiscal policy. Expansionary fiscal policy uses increases in government spending or tax cuts to push the economy out of recession. In an economy with an MPC of 0.75, a $5 billion increase in government spending or a $6.67 billion decrease in personal taxes (producing a $5 billion initial increase in consumption) expands aggregate demand from AD_2 to the downsloping dashed curve. The multiplier then magnifies this initial increase in spending to AD_1. So real GDP rises along the broken horizontal line by $20 billion.

Increased Government Spending Other things equal, a sufficient increase in government spending will shift an economy's aggregate demand curve to the right, from AD_2 to AD_1 in Figure 31.1. To see why, suppose that the recession prompts the government to initiate $5 billion of new spending on highways, education, and health care. We represent this new $5 billion of government spending as the horizontal distance between AD_2 and the dashed downsloping line immediately to its right. At each price level, the amount of real output that is demanded is now $5 billion greater than that demanded before the expansion of government spending.

But the initial increase in aggregate demand is not the end of the story. Through the multiplier effect, the aggregate demand curve shifts to AD_1, a distance that exceeds that represented by the originating $5 billion increase in government purchases. This greater shift occurs because the multiplier process magnifies the initial change in spending into successive rounds of new consumption spending. If the economy's MPC is 0.75, then the simple multiplier is 4. So the aggregate demand curve shifts rightward by four times the distance between AD_2 and the broken downsloping line. Because this *particular* increase in aggregate demand occurs along the horizontal broken-line segment, real output rises by the full extent of the multiplier. Observe that real output rises to $510 billion, up $20 billion from its recessionary level of $490 billion. Concurrently, unemployment falls as firms increase their employment to the full-employment level that existed before the recession.

Tax Reductions Alternatively, the government could reduce taxes to shift the aggregate demand curve rightward, as from AD_2 to AD_1. Suppose the government cuts personal income taxes by $6.67 billion, which increases disposable income by the same amount. Consumption will rise by $5 billion (= MPC of 0.75 × $6.67 billion) and saving will go up by $1.67 billion (= MPS of 0.25 × $6.67 billion). In this case the horizontal distance between AD_2 and the dashed downsloping line in Figure 31.1 represents only the $5 billion initial increase in consumption spending. Again, we call it "initial" consumption spending because the multiplier process yields successive rounds of increased consumption spending. The aggregate demand curve eventually shifts rightward by four times the $5 billion initial increase in consumption produced by the tax cut. Real GDP rises by $20 billion, from $490 billion to $510 billion, implying a multiplier of 4. Employment increases accordingly.

You may have noted that a tax cut must be somewhat larger than the proposed increase in government spending if it is to achieve the same amount of rightward shift in the aggregate demand curve. This is because part of a tax reduction increases saving, rather than consumption. To increase initial consumption by a specific amount, the government must reduce taxes by more than that amount. With an MPC of 0.75, taxes must fall by $6.67 billion for $5 billion of new consumption to be forthcoming because $1.67 billion is saved (not consumed). If the MPC had instead been, say, 0.6, an $8.33 billion reduction in tax collections would have been necessary to increase initial consumption by $5 billion. The smaller the MPC, the greater the tax cut needed to accomplish a specific initial increase in consumption and a specific shift in the aggregate demand curve.

Combined Government Spending Increases and Tax Reductions The government may combine spending increases and tax cuts to produce the desired initial increase in spending and the eventual increase in aggregate demand and real GDP. In the economy depicted in Figure 31.1, the government might increase its spending by $1.25 billion while reducing taxes by $5 billion. As an exercise, you should explain why this combination will produce the targeted $5 billion initial increase in new spending.

If you were assigned Chapter 29, think through these three fiscal policy options in terms of the recessionary-expenditure-gap analysis associated with the aggregate expenditures model (Figure 29.7). And recall from the appendix to Chapter 30 that rightward shifts of the aggregate demand curve relate directly to upward shifts of the aggregate expenditures schedule.

Contractionary Fiscal Policy

When demand-pull inflation occurs, a restrictive or **contractionary fiscal policy** may help control it. This policy consists of government spending reductions, tax increases, or both, designed to decrease aggregate demand and therefore lower or eliminate inflation. Look at Figure 31.2, where the full-employment level of real GDP is $510 billion. The economy starts at equilibrium at point *a*, where the initial aggregate demand curve AD_3 intersects aggregate supply curve AS. Suppose that after going through the multiplier process, a $5 billion initial increase in investment and net export spending shifts the aggregate demand curve to the right by $20 billion, from AD_3 to AD_4. (Ignore the downsloping dashed line for now.) Given the upsloping AS curve, however, the equilibrium GDP does not rise by the full $20 billion. It only rises by $12 billion, to $522 billion, thereby creating an inflationary GDP gap

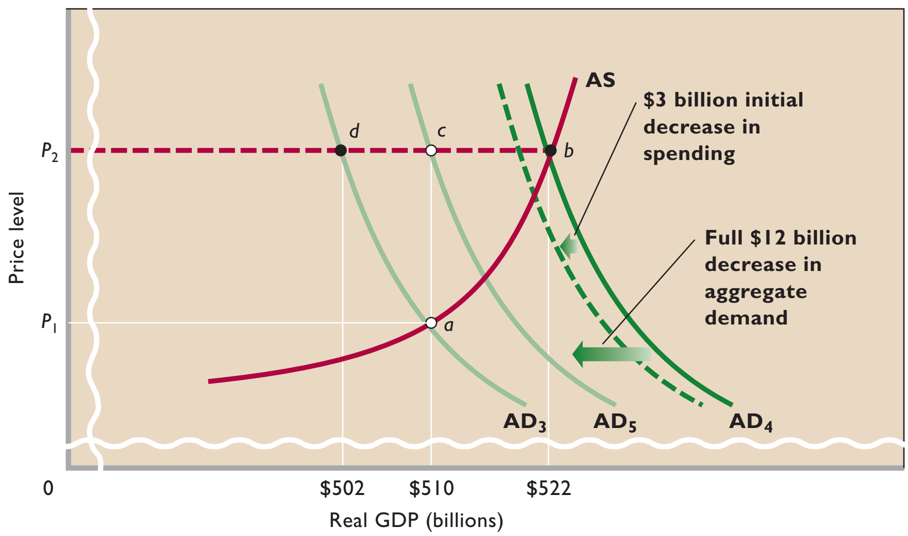


FIGURE 31.2 Contractionary fiscal policy. Contractionary fiscal policy uses decreases in government spending, increases in taxes, or both, to reduce demand-pull inflation. Here, an increase in aggregate demand from AD_3 to AD_4 has driven the economy to point *b* and ratcheted the price level up to P_2, where it becomes inflexible downward. If the economy's MPC is 0.75 and the multiplier therefore is 4, the government can either reduce its spending by $3 billion or increase its taxes by $4 billion (which will decrease consumption by $3 billion) to eliminate the inflationary GDP gap of $12 billion (= $522 billion − $510 billion). Aggregate demand will shift leftward, first from AD_4 to the dashed downsloping curve to its left, and then to AD_5. With the price level remaining at P_2, the economy will move from point *b* to point *c* and the inflationary GDP gap will disappear.

of $12 billion ($522 billion − $510 billion). The upslope of the AS curve means that some of the rightward movement of the AD curve ends up causing demand-pull inflation rather than increased output. As a result, the price level rises from P_1 to P_2 and the equilibrium moves to point *b*.

Without a government response, the inflationary GDP gap will cause further inflation (as input prices rise in the long run to meet the increase in output prices). If the government looks to fiscal policy to eliminate the inflationary GDP gap, its options are the opposite of those used to combat recession. It can (1) decrease government spending, (2) raise taxes, or (3) use some combination of those two policies. When the economy faces demand-pull inflation, fiscal policy should move toward a government **budget surplus**—tax revenues in excess of government spending.

But before discussing how the government can either decrease government spending or increase taxes to move toward a government budget surplus and control inflation, we have to keep in mind that the price level is like a ratchet. While increases in aggregate demand that expand real output beyond the full-employment level tend to ratchet the price level upward, declines in aggregate demand do not seem to push the price level downward. This means that stopping inflation is a matter of halting the rise of the price level, not trying to lower it to the previous level. It also means that the government must take the ratchet effect into account when deciding how big a cut in spending or an increase in taxes it should undertake.

Decreased Government Spending To control demand-pull inflation, the government can decrease aggregate demand by reducing government spending. To see why the ratchet effect matters so much, look at Figure 31.2 and consider what would happen if the government ignored the ratchet effect and attempted to design a spending-reduction policy to eliminate the inflationary GDP gap. Since the $12 billion gap was caused by the $20 billion rightward movement of the aggregate demand curve from AD_3 to AD_4, the government might naively think that it could solve the problem by causing a $20 billion leftward shift of the aggregate demand curve to move it back to where it originally was. It could attempt to do so by reducing government spending by $5 billion and then allowing the multiplier effect to expand that initial decrease into a $20 billion decline in aggregate demand. That would shift the aggregate demand curve leftward by $20 billion, putting it back at AD_3.

This policy would work fine if there were no ratchet effect and if prices were flexible. The economy's equilibrium would move back from point *b* to point *a*, with equilibrium GDP returning to the full-employment level of $510 billion and the price level falling from P_2 back to P_1.

But because there *is* a ratchet effect, this scenario is not what will actually happen. Instead, the ratchet effect implies that the price level is stuck at P_2, so that the broken horizontal line at price level P_2 becomes important to the analysis. The fixed price level means that when the government reduces spending by $5 billion to shift the aggregate demand curve back to AD_3, it will actually cause a recession! The new equilibrium will not be at point *a*. It will be at point *d*, where aggregate demand curve AD_3 crosses the broken horizontal line. At point *d*, real GDP is only $502 billion, $8 billion below the full-employment level of $510 billion.

The problem is that, with the price level downwardly inflexible at P_2, the $20 billion leftward shift of the aggregate demand curve causes a full $20 billion decline in real GDP. None of the change in aggregate demand can be dissipated as a decline in the price level. As a result, equilibrium GDP declines by the full $20 billion, falling from $522 billion to $502 billion and putting it $8 billion below potential output. By not taking the ratchet effect into account, the government has overdone the decrease in government spending, replacing a $12 billion inflationary GDP gap with an $8 billion recessionary GDP gap. This is clearly not what it had in mind.

Here's how it can avoid this scenario. First, the government takes account of the size of the inflationary GDP gap. It is $12 billion. Second, it knows that with the price level fixed, the multiplier will be in full effect. Thus, it knows that any decline in government spending will be multiplied by a factor of 4. It then reasons that government spending will have to decline by only $3 billion rather than $5 billion. Why? Because the $3 billion initial decline in government spending will be multiplied by 4, creating a $12 billion decline in aggregate demand. Under the circumstances, a $3 billion decline in government spending is the correct amount to exactly offset the $12 billion GDP gap. This inflationary GDP gap is the problem that government wants to eliminate. To succeed, it need not undo the full increase in aggregate demand that caused the inflation in the first place.

Graphically, the horizontal distance between AD_4 and the dashed downsloping line to its left represents the $3 billion decrease in government spending. Once the multiplier process is complete, this spending cut will shift the aggregate demand curve leftward from AD_4 to AD_5. With the price level fixed at P_2, the economy will come to equilibrium at point *c*. The economy will operate at its potential output of $510 billion, and the inflationary GDP gap will be eliminated. Furthermore, because the government took the ratchet effect correctly into account, the government will not accidentally push the economy into a recession by making an overly large initial decrease in government spending.

Increased Taxes Just as government can use tax cuts to increase consumption spending, it can use tax *increases* to *reduce* consumption spending. If the economy in Figure 31.2 has an MPC of 0.75, the government must raise taxes by $4 billion to achieve its fiscal policy objective. The $4 billion tax increase reduces saving by $1 billion (= the MPS of 0.25 × $4 billion). This $1 billion reduction in saving, by definition, is not a reduction in spending. But the $4 billion tax increase also reduces consumption spending by $3 billion (= the MPC of 0.75 × $4 billion), as shown by the distance between AD_4 and the dashed downsloping line to its left in Figure 31.2. After the multiplier process is complete, this initial $3 billion decline in consumption will cause aggregate demand to shift leftward by $12 billion at each price level (multiplier of 4 × $3 billion). With the economy moving to point *c*, the inflationary GDP gap will be closed and the inflation will be halted.

Combined Government Spending Decreases and Tax Increases The government may choose to combine spending decreases and tax increases in order to reduce aggregate demand and check inflation. To check your understanding, determine why a $1.5 billion decline in government spending combined with a $2 billion increase in taxes would shift the aggregate demand curve from AD_4 to AD_5. Also, if you were assigned Chapter 29, explain the three fiscal policy options for fighting inflation by referring to the inflationary-expenditure-gap concept developed with the aggregate expenditures model (Figure 29.7). And recall from the appendix to Chapter 30 that leftward shifts of the aggregate demand curve are associated with downshifts of the aggregate expenditures schedule.

Policy Options: *G* or *T*?

Which is preferable as a means of eliminating recession and inflation? The use of government spending or the use of taxes? The answer depends largely on one's view as to whether the government is too large or too small.

Economists who believe there are many unmet social and infrastructure needs usually recommend that government spending be increased during recessions. In times of demand-pull inflation, they usually recommend tax increases. Both actions either expand or preserve the size of government.

Economists who think that the government is too large and inefficient usually advocate tax cuts during recessions and cuts in government spending during times of demand-pull inflation. Both actions either restrain the growth of government or reduce its size.

The point is that discretionary fiscal policy designed to stabilize the economy can be associated with either an expanding government or a contracting government.

QUICK REVIEW 31.1

- Discretionary fiscal policy is the purposeful change of government expenditures and tax collections by government to promote full employment, price stability, and economic growth.
- Expansionary fiscal policy consists of increases in government spending, reductions in taxes, or both, and is designed to expand real GDP by increasing aggregate demand.
- Contractionary fiscal policy entails decreases in government spending, increases in taxes, or both, and is designed to reduce aggregate demand and slow or halt demand-pull inflation.
- To be implemented correctly, contractionary fiscal policy must properly account for the ratchet effect and the fact that the price level will not fall as the government shifts the aggregate demand curve leftward.

Built-In Stability

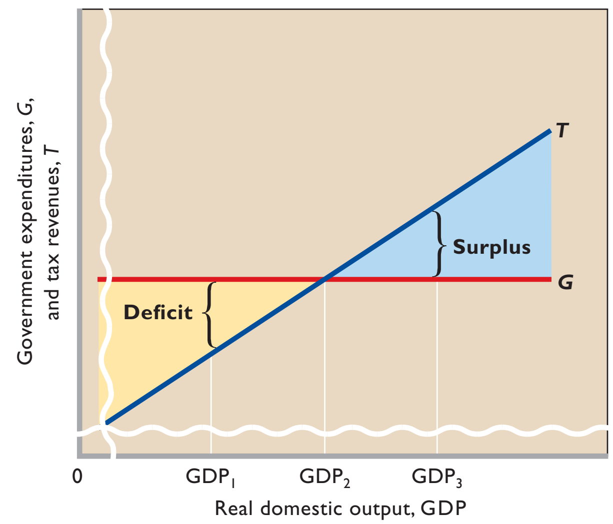
LO31.2 Explain the role of built-in stabilizers in moderating business cycles.

To some degree, government tax revenues change automatically over the course of the business cycle and in ways that stabilize the economy. This automatic response, or built-in stability, constitutes nondiscretionary (or "passive" or "automatic") budgetary policy and results from the makeup of most tax systems. We did not include this built-in stability in our discussion of fiscal policy over the last few pages because we implicitly assumed that the same amount of tax revenue was being collected at each level of GDP. But the actual U.S. tax system is such that *net tax revenues* vary directly with GDP. (Net taxes are tax revenues less transfers and subsidies. From here on, we will use the simpler "taxes" to mean "net taxes.")

Virtually any tax will yield more tax revenue as GDP rises. In particular, personal income taxes have progressive rates and thus generate more-than-proportionate increases in tax revenues as GDP expands. Furthermore, as GDP rises and more goods and services are purchased, revenues from corporate income taxes and from sales taxes and excise taxes also increase. And, similarly, revenues from payroll taxes rise as economic expansion creates more jobs. Conversely, when GDP declines, tax receipts from all these sources also decline.

Transfer payments (or "negative taxes") behave in the opposite way from tax revenues. Unemployment compensation payments and welfare payments decrease during economic expansion and increase during economic contraction.

FIGURE 31.3 Built-in stability. Tax revenues, *T*, vary directly with GDP, and government spending, *G*, is assumed to be independent of GDP. As GDP falls in a recession, deficits occur automatically and help alleviate the recession. As GDP rises during expansion, surpluses occur automatically and help offset possible inflation.



Automatic or Built-In Stabilizers

A **built-in stabilizer** is anything that increases the government's budget deficit (or reduces its budget surplus) during a recession and increases its budget surplus (or reduces its budget deficit) during an expansion without requiring explicit action by policymakers. As Figure 31.3 reveals, this is precisely what the U.S. tax system does. Government expenditures *G* are fixed and assumed to be independent of the level of GDP. Congress decides on a particular level of spending, but it does not determine the magnitude of tax revenues. Instead, it establishes tax rates, and the tax revenues then vary directly with the level of GDP that the economy achieves. Line *T* represents that direct relationship between tax revenues and GDP.

Economic Importance The economic importance of the direct relationship between tax receipts and GDP becomes apparent when we consider that:

- Taxes reduce spending and aggregate demand.
- Reductions in spending are desirable when the economy is moving toward inflation, whereas increases in spending are desirable when the economy is slumping.

As shown in Figure 31.3, tax revenues automatically increase as GDP rises during prosperity, and since taxes reduce household and business spending, they restrain the economic expansion. That is, as the economy moves toward a higher GDP, tax revenues automatically rise and move the budget from deficit toward surplus. In Figure 31.3, observe that the high and perhaps inflationary income level GDP_3 automatically generates a contractionary budget surplus.

Conversely, as GDP falls during recession, tax revenues automatically decline, increasing spending by households and businesses and thus cushioning the economic contraction. With a falling GDP, tax receipts decline and move the government's budget from surplus toward deficit. In Figure 31.3, the low level of income GDP_1 will automatically yield an expansionary budget deficit.

Tax Progressivity Figure 31.3 reveals that the size of the automatic budget deficits or surpluses—and therefore built-in stability—depends on the responsiveness of tax revenues to changes in GDP. If tax revenues change sharply as GDP changes, the slope of line T in the figure will be steep and the vertical distances between T and G (the deficits or surpluses) will be large. If tax revenues change very little when GDP changes, the slope will be gentle and built-in stability will be low.

The steepness of T in Figure 31.3 depends on the tax system itself. In a **progressive tax system,** the average tax rate (= tax revenue/GDP) rises with GDP. In a **proportional tax system,** the average tax rate remains constant as GDP rises. In a **regressive tax system,** the average tax rate falls as GDP rises. The progressive tax system has the steepest tax line T of the three. However, tax revenues will rise with GDP under both the progressive and the proportional tax systems, and they may rise, fall, or stay the same under a regressive tax system. The main point is this: The more progressive the tax system, the greater the economy's built-in stability.

The built-in stability provided by the U.S. tax system has reduced the severity of business fluctuations, perhaps by as much as 8 to 10 percent of the change in GDP that otherwise would have occurred.[1] In recession-year 2009, for example, revenues from the individual income tax fell by a staggering 22 percent. This decline helped keep household spending and real GDP from falling even more than they did. But built-in stabilizers can only dampen, not counteract, swings in real GDP. Discretionary fiscal policy (changes in tax rates and expenditures) or monetary policy (central bank–caused changes in interest rates) therefore may be needed to try to counter a recession or inflation of any appreciable magnitude.

[1]Alan J. Auerbach and Daniel Feenberg, "The Significance of Federal Taxes as Automatic Stabilizers," *Journal of Economic Perspectives*, Summer 2000, p. 54.

Evaluating How Expansionary or Contractionary Fiscal Policy Is Determined

LO31.3 Describe how the cyclically adjusted budget reveals the status of U.S. fiscal policy.

How can we determine whether a government's discretionary fiscal policy is expansionary, neutral, or contractionary? We cannot simply examine the actual budget deficits or surpluses that take place under the current policy because they will necessarily include the automatic changes in tax revenues that accompany every change in GDP. In addition, the expansionary or contractionary strength of any change in discretionary fiscal policy depends not on its absolute size but on how large it is relative to the size of the economy. So, in evaluating the status of fiscal policy, we must adjust deficits and surpluses to eliminate automatic changes in tax revenues and also compare the sizes of the adjusted budget deficits and surpluses to the level of potential GDP.

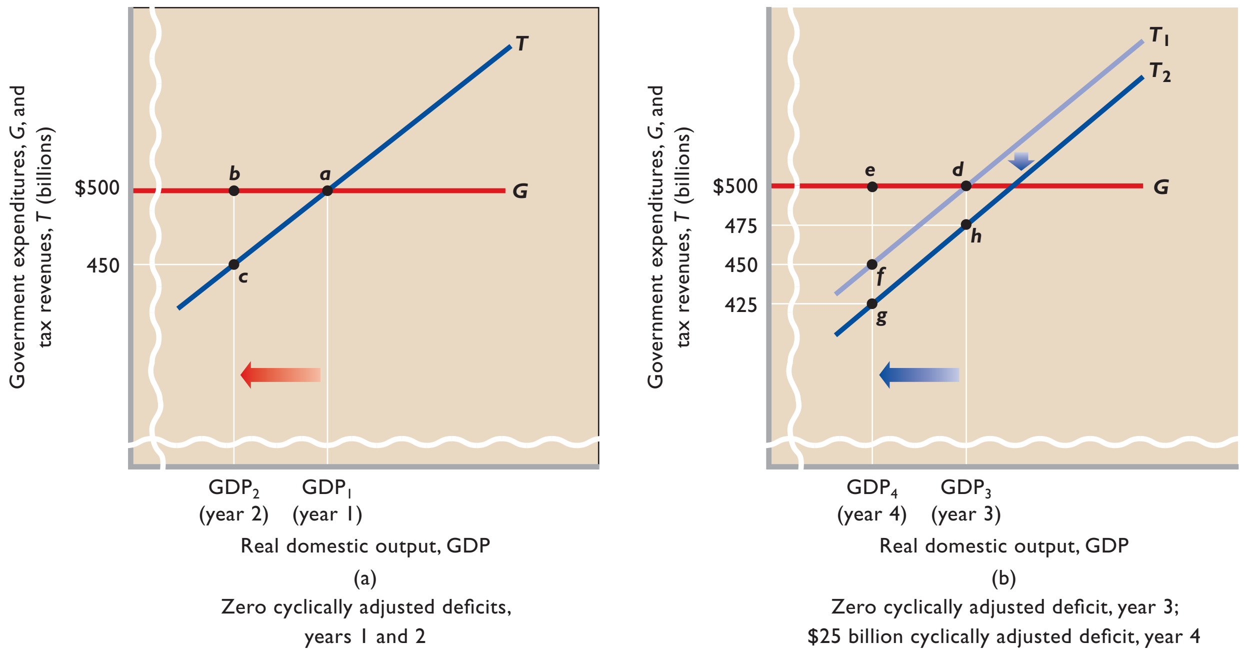
Cyclically Adjusted Budget

Economists use the **cyclically adjusted budget** (also called the *full-employment budget*) to adjust actual federal budget deficits and surpluses to account for the changes in tax revenues that happen automatically whenever GDP changes. The cyclically adjusted budget measures what the federal budget deficit or surplus would have been under existing tax rates and government spending levels if the economy had achieved its full-employment level of GDP (its potential output). The idea essentially is to compare *actual* government expenditures with the tax revenues *that would have occurred* if the economy had achieved full-employment GDP. That procedure removes budget deficits or surpluses that arise simply because of cyclical changes in GDP and thus tell us nothing about whether the government's current discretionary fiscal policy is fundamentally expansionary, contractionary, or neutral.

Consider Figure 31.4a, where line G represents government expenditures and line T represents tax revenues. In full-employment year 1, government expenditures of $500 billion equal tax revenues of $500 billion, as indicated by the intersection of lines G and T at point a. The cyclically adjusted budget deficit in year 1 is zero—government expenditures equal the tax revenues forthcoming at the full-employment output GDP_1. Obviously, the cyclically adjusted deficit *as a percentage of potential GDP* is also zero. The government's fiscal policy is neutral.

Now suppose that a recession occurs and GDP falls from GDP_1 to GDP_2, as shown in Figure 31.4a. Let's also

FIGURE 31.4 Cyclically adjusted deficits. (a) In the left-hand graph, the cyclically adjusted deficit is zero at the full-employment output GDP_1. But it is also zero at the recessionary output GDP_2 because the \$500 billion of government expenditures at GDP_2 equals the \$500 billion of tax revenues that would be forthcoming at the full-employment GDP_1. There has been no change in fiscal policy. (b) In the right-hand graph, discretionary fiscal policy, as reflected in the downward shift of the tax line from T_1 to T_2, has increased the cyclically adjusted budget deficit from zero in year 3 (before the tax cut) to \$25 billion in year 4 (after the tax cut). This is found by comparing the \$500 billion of government spending in year 4 with the \$475 billion of taxes that would accrue at the full-employment GDP_3. Such a rise in the cyclically adjusted deficit (as a percentage of potential GDP) identifies an expansionary fiscal policy.



assume that the government takes no discretionary action, so lines *G* and *T* remain as shown in the figure. Tax revenues automatically fall to \$450 billion (point *c*) at GDP_2, while government spending remains unaltered at \$500 billion (point *b*). A \$50 billion budget deficit (represented by distance *bc*) arises. But this **cyclical deficit** is simply a by-product of the economy's slide into recession, not the result of discretionary fiscal actions by the government. We would be wrong to conclude from this deficit that the government is engaging in an expansionary fiscal policy. The government's fiscal policy has not changed. It is still neutral.

That fact is highlighted when we remove the cyclical part of the deficit and thus consider the cyclically adjusted budget deficit for year 2 in Figure 31.4a. The \$500 billion of government expenditures in year 2 is shown by *b* on line *G*. And, as shown by *a* on line *T*, \$500 billion of tax revenues would have occurred if the economy had achieved its full-employment GDP. Because both *b* and *a* represent \$500 billion, the cyclically adjusted budget deficit in year 2 is zero, as is this deficit as a percentage of potential GDP. Since the cyclically adjusted deficits are zero in both years, we know that government did not change its discretionary fiscal policy, even though a recession occurred and an actual deficit of \$50 billion resulted.

Next, consider Figure 31.4b. Suppose that real output declined from full-employment GDP_3 in year 3 to GDP_4 in year 4. Also suppose that government responded to the recession by reducing tax rates in year 4, as represented by the downward shift of the tax line from T_1 to T_2. What has happened to the size of the cyclically adjusted deficit? Government expenditures in year 4 are \$500 billion, as shown by *e*. Compare that amount with the \$475 billion of tax revenues that would occur if the economy achieved its full-employment GDP. That is, compare position *e* on line *G* with position *h* on line T_2. The \$25 billion of tax revenues by which *e* exceeds *h* is the cyclically adjusted budget deficit for year 4. As a percentage of potential GDP, the cyclically adjusted budget deficit has increased from zero in year 3 (before the tax-rate cut) to some positive percent [= (\$25 billion/GDP_3) × 100] in year 4. This increase in the relative size of the full-employment deficit between the two years reveals that the new fiscal policy is *expansionary*.

In contrast, if we observed a cyclically adjusted deficit (as a percentage of potential GDP) of zero in one year, followed by a cyclically adjusted budget surplus in the next, we could conclude that fiscal policy has changed from being neutral to being contractionary. Because the cyclically adjusted budget adjusts for automatic changes in tax revenues, the increase in the cyclically adjusted budget surplus reveals that government either decreased its spending (G) or increased tax rates such that tax revenues (T) increased. These changes in G and T are precisely the discretionary actions that we have identified as elements of a *contractionary* fiscal policy.

Recent and Projected U.S. Fiscal Policy

LO31.4 Summarize recent U.S. fiscal policy and the projections for U.S. fiscal policy over the next few years.

Table 31.1 lists the actual federal budget deficits and surpluses (column 2) and the cyclically adjusted deficits and surpluses (column 3), as percentages of actual GDP and potential GDP, respectively, between 2000 and 2012. Observe that the cyclically adjusted deficits are generally smaller than the actual deficits. This is because the actual deficits include cyclical deficits, whereas the cyclically adjusted deficits eliminate them. Only cyclically adjusted surpluses and deficits as percentages of potential GDP (column 3) provide the information needed to assess discretionary fiscal policy and determine whether it is expansionary, contractionary, or neutral.

TABLE 31.1 Federal Deficits (−) and Surpluses (+) as Percentages of GDP, 2000–2012

(1) Year	(2) Actual Deficit − or Surplus +	(3) Cyclically adjusted Deficit − or Surplus +*
2000	+2.4	+1.2
2001	+1.3	+0.6
2002	−1.5	−1.2
2003	−3.4	−2.7
2004	−3.5	−3.2
2005	−2.6	−2.6
2006	−1.9	−2.2
2007	−1.2	−1.3
2008	−3.2	−2.9
2009	−10.1	−7.1
2010	−9.0	−5.7
2011	−8.7	−5.6
2012	−7.0	−4.3

*As a percentage of potential GDP.

Source: Congressional Budget Office, **www.cbo.gov**.

Fiscal Policy from 2000 to 2007

Take a look at the data for 2000, for example, which shows that fiscal policy was contractionary that year. Note that the actual budget surplus was 2.4 percent of GDP in 2000 and the cyclically adjusted budget surplus was 1.2 percent of potential GDP. Because the economy was fully employed and corporate profits were strong, tax revenues poured into the federal government and exceeded government expenditures.

But not all was well in 2000. Specifically, the so-called dot-com stock market bubble burst that year, and the U.S. economy noticeably slowed over the latter half of the year. In March 2001 the economy slid into a recession. Congress and the Bush administration responded by cutting taxes by $44 billion in 2001 and scheduling an additional $52 billion of cuts for 2002. These stimulus policies helped boost the economy and offset the recession as well as cushion the economic blow delivered by the September 11, 2001, terrorist attacks. In March 2002 Congress passed further tax cuts totaling $122 billion over two years and extended unemployment benefits.

As Table 31.1 reveals, the cyclically adjusted budget moved from a surplus of 1.2 percent of potential GDP in 2000 to a deficit of −1.2 percent two years later in 2002. Fiscal policy had definitely turned expansionary. Nevertheless, the economy remained sluggish through 2002 and into 2003. In June 2003 Congress again cut taxes, this time by a much larger $350 billion over several years. Specifically, the tax legislation accelerated the reduction of marginal tax rates already scheduled for future years and slashed tax rates on income from dividends and capital gains. It also increased tax breaks for families and small businesses. Note from the table that this tax package increased the cyclically adjusted budget deficit as a percentage of potential GDP to −2.7 percent in 2003. The economy strengthened and both real output and employment grew between 2003 and 2007. By 2007 full employment had been restored, although a −1.3 percent cyclically adjusted budget deficit still remained.

Fiscal Policy during and after the Great Recession

As pointed out in previous chapters, major economic trouble began in 2007. In the summer of 2007, a crisis in the market for mortgage loans flared up. Later in 2007 that crisis spread rapidly to other financial markets, threatened the survival of several major U.S. financial institutions, and severely disrupted the entire financial system. As credit markets began to freeze, general pessimism spread beyond the financial markets to the overall economy. Businesses and households retrenched on their borrowing and spending, and in December 2007 the economy entered a recession.

Over the following two years, it became known as the Great Recession—one of the steepest and longest economic downturns since the 1930s.

In 2008 Congress acted rapidly to pass an economic stimulus package. This law provided a total of $152 billion in stimulus, with some of it coming as tax breaks for businesses, but most of it delivered as checks of up to $600 each to taxpayers, veterans, and Social Security recipients.

As a percentage of GDP, the *actual* federal budget deficit jumped from −1.2 percent in 2007 to −3.2 percent in 2008. This increase resulted from an automatic drop-off of tax revenues during the recession, along with the tax rebates (fiscal stimulus checks) paid out in 2008. As shown in Table 31.1, the *cyclically adjusted* budget deficit rose from −1.3 percent of potential GDP in 2007 to −2.9 percent in 2008. This increase in the cyclically adjusted budget reveals that fiscal policy in 2008 was expansionary.

The government hoped that those receiving checks would spend the money and thus boost consumption and aggregate demand. But households instead saved substantial parts of the money from the checks or used some of the money to pay down credit card loans. Although this stimulus plan boosted output somewhat in mid-2008, it was neither as expansionary nor long-lasting as policymakers had hoped. The continuing forces of the Great Recession simply overwhelmed the policy.

With the economy continuing its precipitous slide, the Obama administration and Congress enacted the American Recovery and Reinvestment Act of 2009. This gigantic $787 billion program—coming on top of a $700 billion rescue package for financial institutions—consisted of low- and middle-income tax rebates, plus large increases in expenditures on infrastructure, education, and health care. The idea was to flood the economy with additional spending to try to boost aggregate demand and get people back to work.

The tax cuts in the package were aimed at lower- and middle-income individuals and households, who were thought to be more likely than high-income people to spend (rather than save) the extra income from the tax rebates. Rather than sending out lump-sum stimulus checks as in 2008, the new tax rebates showed up as small increases in workers' monthly payroll checks. With smaller amounts per month rather than a single large check, the government hoped that people would spend the bulk of their enhanced income—rather than save it as they had done with the one-time-only, lump-sum checks received in 2008. The second part of the fiscal policy (60 percent of the funding) consisted of increases in government expenditures on a wide assortment of programs, including transportation, education, and aid to state governments. The highly stimulative fiscal policy for 2009 is fully reflected in column 3 of Table 31.1. The cyclically adjusted budget deficit rose dramatically from −2.9 percent of potential GDP in 2008 to a very high −7.1 percent of potential GDP in 2009.

The recession officially ended during the summer of 2009, but the economy did not rebound vigorously. Unemployment remained elevated and tax collections were low due to a stagnant economy. As a result, policymakers decided to continue with large amounts of fiscal stimulus. Annual actual (not cyclically adjusted) budget deficits amounted to −9.0, −8.7, and −7.0 percent of GDP, respectively, in years 2010, 2011, and 2012. The cyclically adjusted budget deficits for those years were, respectively, −5.7, −5.6, and −4.3 percent of potential GDP. Thus, while the intensity of fiscal stimulus was gradually diminishing, it remained very high by historical standards. The recession had been exceptionally strong, and so was the ongoing fiscal response.

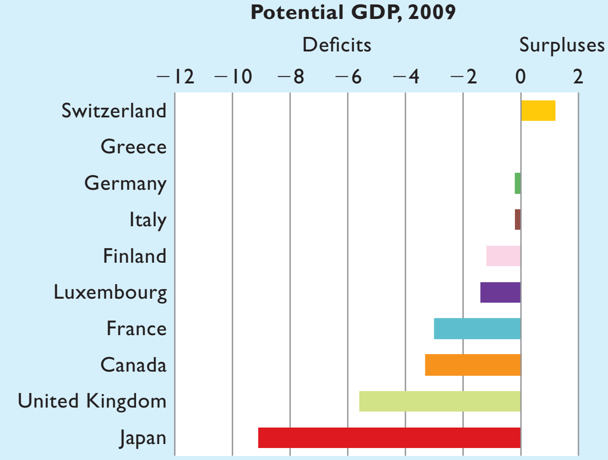
Other nations also experienced recessions and also responded with expansionary fiscal policies. Global Perspective 31.1 shows the magnitudes of the cyclically

GLOBAL PERSPECTIVE 31.1

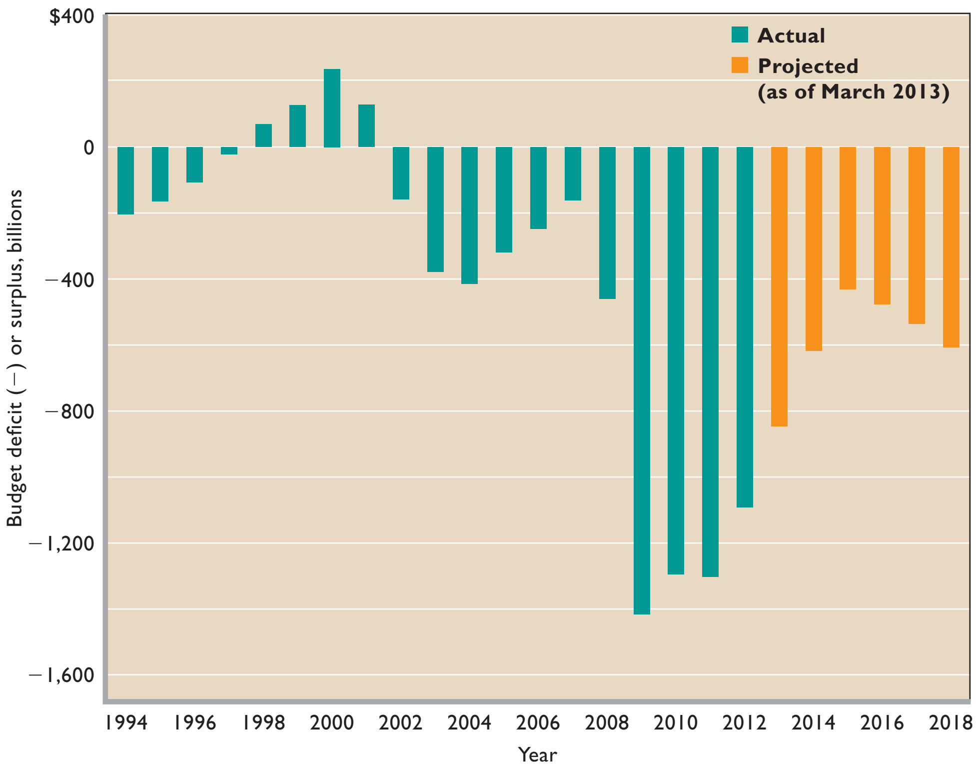
Cyclically Adjusted Budget Deficits or Surpluses as a Percentage of Potential GDP, Selected Nations

Because of the global recession, in 2009 all but a few of the world's major nations had cyclically adjusted budget deficits. These deficits varied as percentages of potential GDP, but they each reflected some degree of expansionary fiscal policy.

Cyclically Adjusted Budget Surplus or Deficit as a Percentage of Potential GDP, 2009



Source: OECD Economic Outlook, **http://www.oecd.org/eco/economicoutlook.htm**.



Source: Congressional Budget Office, **www.cbo.gov**.

FIGURE 31.5 Federal budget deficits and surpluses, actual and projected, fiscal years 1994–2018 (in billions of nominal dollars). The annual budget deficits of 1994 through 1997 gave way to budget surpluses from 1998 through 2001. Deficits reappeared in 2002 and declined through 2007. They greatly ballooned in recessionary years 2008 and 2009 and are projected to remain high for many years to come.

adjusted surpluses and deficits of a number of countries in 2009.

Past and Projected Budget Deficits and Surpluses

Figure 31.5 shows the absolute magnitudes of actual (not cyclically adjusted) U.S. budget surpluses and deficits, here from 1994 through 2012. It also shows the projected future deficits through 2018, as estimated by the Congressional Budget Office (CBO). In recession year 2009, the federal budget deficit reached $1,413 billion, mainly but not totally due to reduced tax revenues from lower income and record amounts of stimulus spending. The CBO projects high deficits for several years to come. But projected deficits and surpluses are subject to large and frequent changes, as government alters its fiscal policy and GDP growth accelerates or slows. So we suggest that you update this figure by going to the Congressional Budget Office Web site, **www.cbo.gov**, and find the document titled *The Budget and Economic Outlook*. Near the start of that document, you should find Summary Table 1. The numbers are in the row labeled "Deficit (−) or Surplus."

Problems, Criticisms, and Complications of Implementing Fiscal Policy

LO31.5 Discuss the problems that governments may encounter in enacting and applying fiscal policy.

Economists recognize that governments may encounter a number of significant problems in enacting and applying fiscal policy.

Problems of Timing

Several problems of timing may arise in connection with fiscal policy:

- ***Recognition lag*** The recognition lag is the time between the beginning of recession or inflation and the certain awareness that it is actually happening. This lag arises because the economy does not move smoothly through the business cycle. Even during good times, the economy has slow months interspersed with months of rapid growth and expansion. This makes recognizing a recession difficult since several slow months will have to happen in succession before people can conclude with any confidence that the good times are over and a recession has begun.

The same is true with inflation. Even periods of moderate inflation have months of high inflation—so that several high-inflation months must come in sequence before people can confidently conclude that inflation has moved to a higher level.

Attempts to reduce the length of the recognition lag by trying to predict the future course of the economy also have proven to be highly difficult, at best. As a result, the economy is often 4 to 6 months into a recession or inflation before the situation is clearly discernible in the relevant statistics. Due to this recognition lag, the economic downslide or the inflation may become more serious than it would have if the situation had been identified and acted on sooner.

- ***Administrative lag*** The wheels of democratic government turn slowly. There will typically be a significant lag between the time the need for fiscal action is recognized and the time action is taken. Following the terrorist attacks of September 11, 2001, the U.S. Congress was stalemated for 5 months before passing a compromise economic stimulus law in March 2002. (In contrast, the Federal Reserve began lowering interest rates the week after the attacks.)
- ***Operational lag*** A lag also occurs between the time fiscal action is taken and the time that action affects output, employment, or the price level. Although changes in tax rates can be put into effect relatively quickly once new laws are passed, government spending on public works—new dams, interstate highways, and so on—requires long planning periods and even longer periods of construction. Such spending is of questionable use in offsetting short (for example, 6- to 12-month) periods of recession. Consequently, discretionary fiscal policy has increasingly relied on tax changes rather than on changes in spending as its main tool.

Political Considerations

Fiscal policy is conducted in a political arena. That reality not only may slow the enactment of fiscal policy but also may create the potential for political considerations swamping economic considerations in its formulation. It is a human trait to rationalize actions and policies that are in one's self-interest. Politicians are very human—they want to get reelected. A strong economy at election time will certainly help them. So they may favor large tax cuts under the guise of expansionary fiscal policy even though that policy is economically inappropriate. Similarly, they may rationalize increased government spending on popular items such as farm subsidies, health care, highways, education, and homeland security.

At the extreme, elected officials and political parties might collectively "hijack" fiscal policy for political purposes, cause inappropriate changes in aggregate demand, and thereby cause (rather than avert) economic fluctuations. For instance, before an election, they may try to stimulate the economy to improve their reelection hopes. And then after the election, they may try to use contractionary fiscal policy to dampen the excessive aggregate demand that they caused with their preelection stimulus. In short, elected officials may cause so-called **political business cycles**—swings in overall economic activity and real GDP resulting from election-motivated fiscal policy, rather than from inherent instability in the private sector. Political business cycles are difficult to document and prove, but there is little doubt that political considerations weigh heavily in the formulation of fiscal policy. The question is how often those political considerations run counter to "sound economics."

Future Policy Reversals

Fiscal policy may fail to achieve its intended objectives if households expect future reversals of policy. Consider a tax cut, for example. If taxpayers believe the tax reduction is temporary, they may save a large portion of their tax cut, reasoning that rates will return to their previous level in the future. They save more now so that they will be able to draw on this extra savings to maintain their future consumption levels if taxes do indeed rise again in the future. So a tax reduction thought to be temporary may not increase present consumption spending and aggregate demand by as much as our simple model (Figure 31.1) suggests.

The opposite may be true for a tax increase. If taxpayers think it is temporary, they may reduce their saving to pay the tax while maintaining their present consumption. They may reason they can restore their saving when the tax rate again falls. So the tax increase may not reduce current consumption and aggregate demand by as much as policymakers intended.

To the extent that this so-called *consumption smoothing* occurs over time, fiscal policy will lose some of its strength. The lesson is that tax-rate changes that households view as permanent are more likely to alter consumption and aggregate demand than tax changes they view as temporary.

Offsetting State and Local Finance

The fiscal policies of state and local governments are frequently *pro-cyclical*, meaning that they worsen rather than correct recession or inflation. Unlike the federal government, most state and local governments face constitutional or other legal requirements to balance their budgets. Like households and private businesses, state and local governments increase their expenditures during prosperity and cut them during recession.

During the Great Depression of the 1930s, most of the increase in federal spending was offset by decreases in state and local spending. During and immediately following the recession of 2001, many state and local governments had to offset lower tax revenues resulting from the reduced personal income and spending of their citizens. They offset the decline in revenues by raising tax rates, imposing new taxes, and reducing spending.

In view of these past experiences, the $787 billion fiscal package of 2009 made a special effort to reduce this problem by giving substantial aid dollars to state governments. Because of the sizable federal aid, the states did not have to increase taxes and reduce expenditure by as much as otherwise. So their collective fiscal actions did not fight as much against the increase in aggregate demand that the federal government wanted to achieve with its tax cuts and expenditure increases.

Crowding-Out Effect

Another potential flaw of fiscal policy is the so-called **crowding-out effect:** An expansionary fiscal policy (deficit spending) may increase the interest rate and reduce investment spending, thereby weakening or canceling the stimulus of the expansionary policy. The rising interest rate might also potentially crowd out interest-sensitive consumption spending (such as purchasing automobiles on credit). But since investment is the most volatile component of GDP, the crowding-out effect focuses its attention on investment and whether the stimulus provided by deficit spending may be partly or even fully neutralized by an offsetting reduction in investment spending.

To see the potential problem, realize that whenever the government borrows money (as it must if it is deficit spending), it increases the overall demand for money. If the monetary authorities are holding the money supply constant, this increase in demand will raise the price paid for borrowing money: the interest rate. Because investment spending varies inversely with the interest rate, some investment will be choked off or "crowded out."

ORIGIN OF THE IDEA

O31.2
Crowding out

Economists vary in their opinions about the strength of the crowding-out effect. An important thing to keep in mind is that crowding out is likely to be less of a problem when the economy is in recession because investment demand tends to be weak. Why? Because output purchases slow during recessions and therefore most businesses end up with substantial excess capacity. As a result, they do not have much incentive to add new machinery or build new factories. After all, why should they add capacity when some of the capacity they already have is lying idle?

With investment demand weak during a recession, the crowding-out effect is likely to be very small. Simply put, there is not much investment for the government to crowd out. Even if deficit spending does increase the interest rate, the effect on investment may be fully offset by the improved investment prospects that businesses expect from the fiscal stimulus.

By contrast, when the economy is operating at or near full capacity, investment demand is likely to be quite strong so that crowding out will probably be a much more serious problem. When the economy is booming, factories will be running at or near full capacity and firms will have high investment demand for two reasons. First, equipment running at full capacity wears out fast, so firms will be investing substantial amounts just to replace machinery and equipment that wears out and depreciates. Second, the economy is likely to be growing overall so that firms will be heavily investing to *add* to their production capacity to take advantage of the greater anticipated demand for their outputs.

Current Thinking on Fiscal Policy

Where do these complications leave us as to the advisability and effectiveness of discretionary fiscal policy? In view of the complications and uncertain outcomes of fiscal policy, some economists argue that it is better not to engage in it at all. Those holding that view point to the superiority of monetary policy (changes in interest rates engineered by the Federal Reserve) as a stabilizing device or believe that most economic fluctuations tend to be mild and self-correcting.

But most economists believe that fiscal policy remains an important, useful policy lever in the government's macroeconomic toolkit. The current popular view is that fiscal policy can help push the economy in a particular direction but cannot fine-tune it to a precise macroeconomic outcome. Mainstream economists generally agree that monetary policy is the best month-to-month stabilization tool for the U.S. economy. If monetary policy is doing its job, the government should maintain a relatively neutral fiscal policy, with a cyclically adjusted budget deficit or surplus of no more than 2 percent of potential GDP. It should hold major discretionary fiscal policy in reserve to help counter situations where recession threatens to be deep and long lasting, as in 2008 and 2009, or where a substantial reduction in aggregate demand might help the Federal Reserve to quell a major bout of inflation.

Finally, economists agree that any proposed fiscal policy should be evaluated for its potential positive and negative impacts on long-run productivity growth. The short-run policy tools used for conducting active fiscal policy often have long-run impacts. Countercyclical fiscal policy should be shaped to strengthen, or at least not impede, the growth of long-run aggregate supply (shown as a rightward shift of the long-run aggregate supply curve in Figure 30.5). For example, a tax cut might be structured to enhance work effort, strengthen investment, and encourage innovation. Alternatively, an increase in government spending might center on preplanned projects for public capital (highways, mass transit, ports, airports), which are complementary to private investment and thus support long-term economic growth.

QUICK REVIEW 31.2

- Automatic changes in net taxes (taxes minus transfers) add a degree of built-in stability to the economy.
- Cyclical deficits arise from declines in net tax revenues that automatically occur as the economy recedes and incomes and profits fall.
- The cyclically adjusted budget eliminates cyclical effects on net tax revenues; it compares actual levels of government spending to the projected levels of net taxes that would occur if the economy were achieving its full-employment output.
- Time lags, political problems, expectations, and state and local finances complicate fiscal policy.
- The crowding-out effect indicates that an expansionary fiscal policy may increase the interest rate and reduce investment spending.

The U.S. Public Debt

LO31.6 Discuss the size, composition, and consequences of the U.S. public debt.

The U.S. national debt, or **public debt,** is essentially the accumulation of all past federal deficits and surpluses. The deficits have greatly exceeded the surpluses and have emerged mainly from war financing, recessions, and fiscal policy. In 2012 the total public debt was $16.4 trillion—$11.6 trillion held by the public, excluding the Federal Reserve; and $4.8 trillion held by federal agencies and the Federal Reserve. Between 2007 and 2009, the total public debt expanded by a huge $2.9 trillion. During the Great Recession, federal tax revenues plummeted because incomes and profit fell, and federal expenditures jumped because of huge spending to rescue failing financial institutions and to stimulate the shrinking economy.

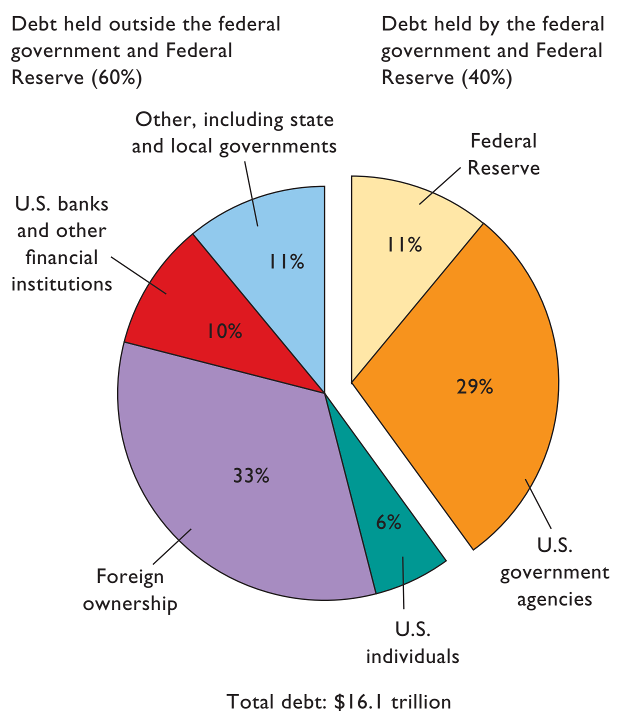
Because large annual deficits continued over the next several years, the total public debt grew past $17.0 trillion in late 2013. It had doubled in just seven years, growing from $8.5 trillion in late 2006 to $17.0 trillion in late 2013. You can find the current size of the public debt at the Web site of the Department of Treasury, Bureau of the Public Debt, at **www.treasurydirect.gov/NP/BPDLogin?application=np**. At this site, you will see that the U.S. Treasury defines "the public" to include the Federal Reserve. But because the Federal Reserve is the nation's central bank, economists view it as essentially part of the federal government and not part of the public. Economists typically focus on the part of the debt that is not owned by the federal government and the Federal Reserve.

Ownership

The total public debt of $16.4 trillion represents the total amount of money owed by the federal government to the holders of **U.S. government securities:** financial instruments issued by the federal government to borrow money to finance expenditures that exceed tax revenues. U.S. government securities (loan instruments) are of four types: *Treasury bills* (short-term securities), *Treasury notes* (medium-term securities), *Treasury bonds* (long-term securities), and *U.S. savings bonds* (long-term, nonmarketable bonds).

Figure 31.6 shows that the public, sans the Federal Reserve, held 60 percent of the federal debt in 2012 and that federal government agencies and the Federal Reserve held the remaining 40 percent. The federal agencies hold U.S. government securities as risk-free assets that they can cash in as needed to make latter payments. The Federal Reserve holds these securities to facilitate the "open-market operations" that it uses to control the nation's money supply (Chapter 34). Observe that "the public" in the pie chart consists of individuals here and abroad, state and local governments, and U.S. financial institutions. Foreigners held about 33 percent of the total U.S. public debt in 2012, meaning that most of the U.S. public debt is held internally and not externally. Americans owed 67 percent of the public debt to Americans. Of the $5.3 trillion

FIGURE 31.6 Ownership of the total public debt, 2012. The $16.1 trillion public debt can be divided into the proportion held by the public, excluding the Federal Reserve (60 percent), and the proportion held by federal agencies and the Federal Reserve System (40 percent). Of the total debt, 33 percent is foreign-owned.



Source: Economic Report of the President, 2013, **www.gpo.gov/fdsys/browse/collection.action?collectionCode=ERP**; authors' derivation from Table B-89, September 2012 data. Federal Reserve percentage is from the U.S. Treasury, **www.fms.treas.gov/bulletin**.

of debt held by foreigners, China held 22 percent, Japan held 20 percent, and oil-exporting nations held 5 percent.

Debt and GDP

A simple statement of the absolute size of the debt ignores the fact that the wealth and productive ability of the U.S. economy is also vast. A wealthy, highly productive nation can incur and carry a large public debt much more easily than a poor nation can. A more meaningful measure of the public debt relates it to an economy's GDP. Figure 31.7 shows the yearly relative sizes of the U.S. public debt held outside the Federal Reserve and federal agencies. In 2012 the percentage was 70 percent. Most noticeably, the percentage rose dramatically starting in 2008 because of massive annual budget deficits.

International Comparisons

It is not uncommon for countries to have sizable public debts. As shown in Global Perspective 31.2, the public debt as a percentage of real GDP in the United States is neither particularly high nor low relative to such debt percentages in other advanced industrial nations.

Interest Charges

Many economists conclude that the primary burden of the debt is the annual interest charge accruing on the bonds sold to finance the debt. In 2012 interest on the total public debt was $360 billion. Although this amount is sizable in absolute terms, it was only 2.3 percent of GDP for

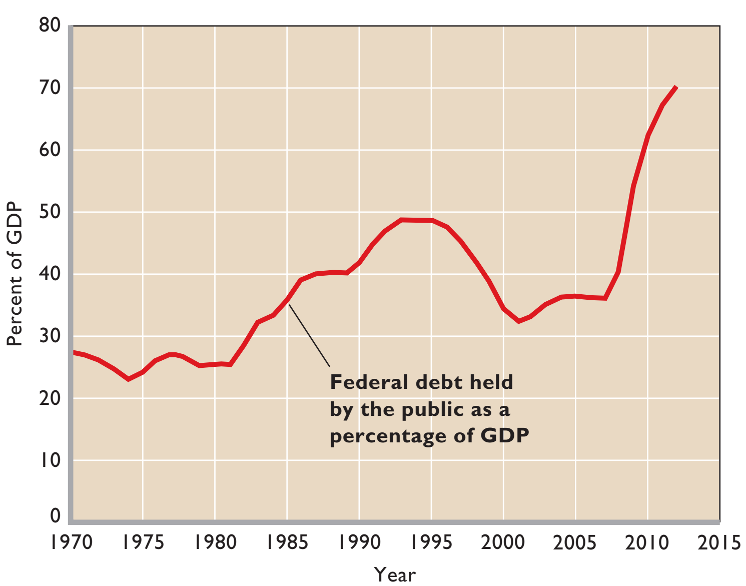


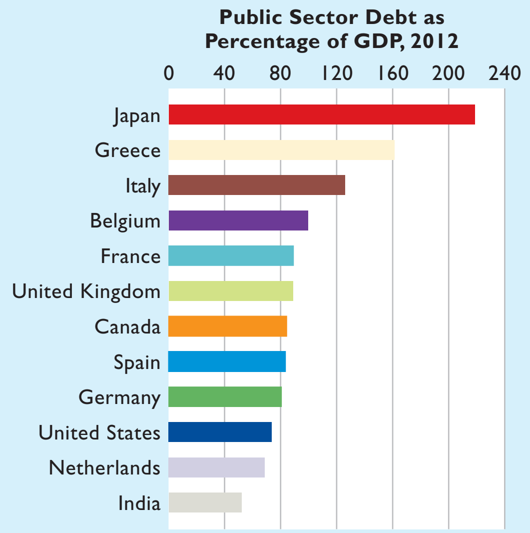
FIGURE 31.7 Federal debt held by the public, excluding the Federal Reserve, as a percentage of GDP, 1970–2012. As a percentage of GDP, the federal debt held by the public (held outside the Federal Reserve and federal government agencies) increased sharply over the 1980–1995 period and declined significantly between 1995 and 2001. Since 2001, the percentage has gone up again, and jumped abruptly and sharply starting in 2008.

Source: FRED II database, St. Louis Federal Reserve Bank, **research.stlouisfed.org/fred2/**.

GLOBAL PERSPECTIVE 31.2

Publicly Held Debt: International Comparisons

Although the United States has the world's largest public debt, a number of other nations have larger debts as percentages of their GDPs.



Source: CIA World Factbook, **www.cia.gov/library/publications/the-world-factbook/**. These debt calculations encompass federal, state, and local debt, including the debt of government-owned enterprises (not just federal debt as in Figure 31.7) and are 2012 estimates.

2012. So, the federal government had to collect taxes equal to 2.3 percent of GDP to service the total public debt. This percentage was unchanged from 2000 despite the much higher total public debt. That was possible because interest rates were being kept extremely low by the Federal Reserve in order to help stimulate the economy in the wake of the Great Recession (Chapter 34).

False Concerns

You may wonder if the large public debt might bankrupt the United States or at least place a tremendous burden on your children and grandchildren. Fortunately, these are largely false concerns. People were wondering the same things 50 years ago!

Bankruptcy

The large U.S. public debt does not threaten to bankrupt the federal government, leaving it unable to meet its financial obligations. There are two main reasons: refinancing and taxation.

Refinancing As long as the U.S. public debt is viewed by lenders as manageable and sustainable, the public debt is easily refinanced. As portions of the debt come due on maturing Treasury bills, notes, and bonds each month, the government does not cut expenditures or raise taxes to provide the funds required. Rather, it refinances the debt by selling new bonds and using the proceeds to pay holders of the maturing bonds. The new bonds are in strong demand because lenders can obtain a market-determined interest return with no risk of default by the federal government.

Of course, refinancing could become an issue with a high enough debt-to-GDP ratio. Some countries such as Greece have run into this problem. High and rising ratios in the United States might raise fears that the U.S. government might be unable to pay back loans as they come due. But, with the present U.S. debt-to-GDP ratio and the prospects of long-term economic growth, this is a false concern for the United States.

Taxation The federal government has the constitutional authority to levy and collect taxes. A tax increase is a government option for gaining sufficient revenue to pay interest and principal on the public debt. Financially distressed private households and corporations cannot extract themselves from their financial difficulties by taxing the public. If their incomes or sales revenues fall short of their expenses, they can indeed go bankrupt. But the federal government does have the option to impose new taxes or increase existing tax rates if necessary to finance its debt. Such tax hikes may be politically unpopular and may weaken incentives to work and invest, but they *are* a means of raising funds to finance the debt.

Burdening Future Generations

In 2012 public debt per capita was $52,396. Was each child born in 2012 handed a $52,396 bill from the federal government? Not really. The public debt does not impose as much of a burden on future generations as commonly thought.

The United States owes a substantial portion of the public debt to itself. U.S. citizens and institutions (banks, businesses, insurance companies, governmental agencies, and trust funds) own about 67 percent of the U.S. government securities. Although that part of the public debt is a liability to Americans (as taxpayers), it is simultaneously an asset to Americans (as holders of Treasury bills, Treasury notes, Treasury bonds, and U.S. savings bonds).

To eliminate the American-owned part of the public debt would require a gigantic transfer payment from Americans to Americans. Taxpayers would pay higher taxes, and holders of the debt would receive an equal amount for their U.S. government securities. Purchasing power in the

United States would not change. Only the repayment of the 33 percent of the public debt owned by foreigners would negatively impact U.S. purchasing power.

The public debt increased sharply during the Second World War. But the decision to finance military purchases through the sale of government bonds did not shift the economic burden of the war to future generations. The economic cost of the Second World War consisted of the civilian goods society had to forgo in shifting scarce resources to war goods production (recall production possibilities analysis). Regardless of whether society financed this reallocation through higher taxes or through borrowing, the real economic burden of the war would have been the same. That burden was borne almost entirely by those who lived during the war. They were the ones who did without a multitude of consumer goods to enable the United States to arm itself and its allies.

The next generation inherited the debt from the war but also an equal amount of government bonds that would pay them cash in future years. It also inherited the enormous benefits from the victory—namely, preserved political and economic systems at home and the "export" of those systems to Germany, Italy, and Japan. Those outcomes enhanced postwar U.S. economic growth and helped raise the standard of living of future generations of Americans.

Substantive Issues

Although the preceding issues relating to the public debt are false concerns, a number of substantive issues are not. Economists, however, attach varying degrees of importance to them.

Income Distribution

The distribution of ownership of government securities is highly uneven. Some people own much more than the $52,396-per-person portion of government securities; other people own less or none at all. In general, the ownership of the public debt is concentrated among wealthier groups, who own a large percentage of all stocks and bonds. Because the overall federal tax system is only slightly progressive, payment of interest on the public debt mildly increases income inequality. Income is transferred from people who, on average, have lower incomes to the higher-income bondholders. If greater income equality is one of society's goals, then this redistribution is undesirable.

Incentives

The current public debt necessitates annual interest payments of $360 billion. With no increase in the size of the debt, that interest charge must be paid out of tax revenues. Higher taxes may dampen incentives to bear risk, to innovate, to invest, and to work. So, in this indirect way, a large public debt may impair economic growth and therefore impose a burden of reduced output (and income) on future generations.

Foreign-Owned Public Debt

The 33 percent of the U.S. debt held by citizens and institutions of foreign countries *is* an economic burden to Americans. Because we do not owe that portion of the debt "to ourselves," the payment of interest and principal on this **external public debt** enables foreigners to buy some of our output. In return for the benefits derived from the borrowed funds, the United States transfers goods and services to foreign lenders. Of course, Americans also own debt issued by foreign governments, so payment of principal and interest by those governments transfers some of their goods and services to Americans.

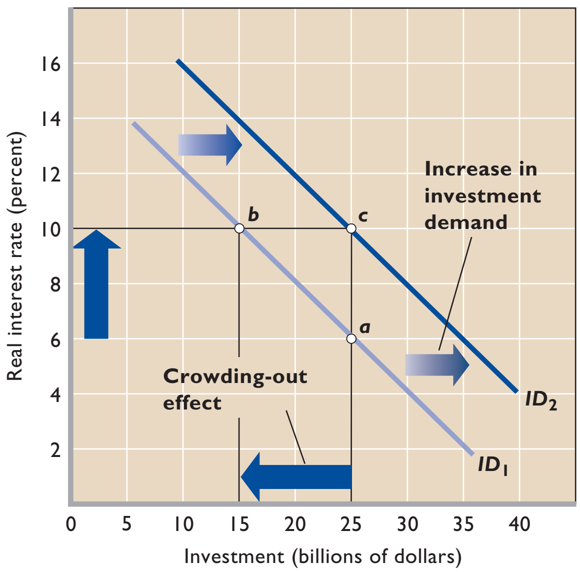
Crowding-Out Effect Revisited

A potentially more serious problem is the financing (and continual refinancing) of the large public debt, which can transfer a real economic burden to future generations by passing on to them a smaller stock of capital goods. This possibility involves the previously discussed crowding-out effect: the idea that public borrowing drives up real interest rates, which reduces private investment spending. If public borrowing only happened during recessions, crowding out would not likely be much of a problem. Because private investment demand tends to be weak during recessions, any increase in interest rates caused by public borrowing will at most cause a small reduction in investment spending.

In contrast, the need to continuously finance a large public debt may be more troublesome. At times, that financing requires borrowing large amounts of money when the economy is near or at its full-employment output. Because this usually is when private demand is strong, any increase in interest rates caused by the borrowing necessary to refinance the debt may result in a substantial decline in investment spending. If the amount of current investment crowded out is extensive, future generations will inherit an economy with a smaller production capacity and, other things equal, a lower standard of living.

A Graphical Look at Crowding Out We know from Chapter 28 that the amount of investment spending is

FIGURE 31.8 The investment demand curve and the crowding-out effect. If the investment demand curve (ID_1) is fixed, the increase in the interest rate from 6 percent to 10 percent caused by financing a large public debt will move the economy from *a* to *b*, crowding out $10 billion of private investment and decreasing the size of the capital stock inherited by future generations. However, if the public goods enabled by the debt improve the investment prospects of businesses, the private investment demand curve will shift rightward, as from ID_1 to ID_2. That shift may offset the crowding-out effect wholly or in part. In this case, it moves the economy from *a* to *c*.



inversely related to the real interest rate. When graphed, that relationship is shown as a downsloping investment demand curve, such as either ID_1 or ID_2 in Figure 31.8. Let's first consider curve ID_1. (Ignore curve ID_2 for now.) Suppose that government borrowing increases the real interest rate from 6 percent to 10 percent. Investment spending will then fall from $25 billion to $15 billion, as shown by the economy's move from *a* to *b*. That is, the financing of the debt will compete with the financing of private investment projects and crowd out $10 billion of private investment. So the stock of private capital handed down to future generations will be $10 billion less than it would have been without the need to finance the public debt.

Public Investments and Public-Private Complementarities But even with crowding out, two factors could partly or fully offset the net economic burden shifted to future generations. First, just as private expenditures may involve either consumption or investment, so it is with public goods. Part of the government spending enabled by the public debt is for public investment outlays (for example, highways, mass transit systems, and electric power facilities) and "human capital" (for example, investments in education, job training, and health). Like private expenditures on machinery and equipment, those **public investments** increase the economy's future production capacity. Because of the financing through debt, the stock of public capital passed on to future generations may be higher than otherwise. That greater stock of public capital may offset the diminished stock of private capital resulting from the crowding-out effect, leaving overall production capacity unimpaired.

So-called public-private complementarities are a second factor that could reduce the crowding-out effect. Some public and private investments are complementary. Thus, the public investment financed through debt could spur some private-sector investment by increasing its expected rate of return. For example, a federal building in a city may encourage private investment in the form of nearby office buildings, shops, and restaurants. Through its complementary effect, the spending on public capital may shift the private investment demand curve to the right, as from ID_1 to ID_2 in Figure 31.8. Even though the government borrowing boosts the interest rate from 6 percent to 10 percent, total private investment need not fall. In the case shown as the move from *a* to *c* in Figure 31.8, it remains at $25 billion. Of course, the increase in investment demand might be smaller than that shown. If it were smaller, the crowding-out effect would not be fully offset. But the point is that an increase in investment demand may counter the decline in investment that would otherwise result from the higher interest rate.

QUICK REVIEW 31.3

- The U.S. public debt—$16.4 trillion in 2012—is essentially the total accumulation of all past federal budget deficits and surpluses; about 33 percent of the U.S. public debt is held by foreigners.
- The U.S. public debt held by the public (excluding the Federal Reserve) was 70 percent of GDP in 2012, up from 30 percent in 2000.
- The federal government is in no danger of going bankrupt because it needs only to refinance (not retire) the public debt and it can raise revenues, if needed, through higher taxes.
- The borrowing and interest payments associated with the public debt may (a) increase income inequality; (b) require higher taxes, which may dampen incentives; and (c) impede the growth of the nation's stock of capital through crowding out of private investment.

LAST WORD

The Social Security and Medicare Shortfalls

Social Security and Medicare Face Gigantic Future Funding Shortfalls. Metaphorically Speaking, Some Economists See These Programs as Financial and Political Time Bombs.

The American population, on average, is getting decidedly older. The percentage of the population age 62 or older will rise substantially over the next several decades, with the greatest increases for people age 75 and above. In the future, more people will be receiving Social Security benefits (income during retirement) and Medicare (medical care during retirement) for longer periods. Each person's benefits will be paid for by fewer workers. The number of workers per Social Security and Medicare beneficiary was roughly 5:1 in 1960. Today it is 3:1, and by 2040 it will be only 2:1.

The combined cost of the Social Security and Medicare programs was 8.5 percent of GDP in 2011, and that percentage is projected to grow to 12 percent of GDP in 2035 and 13 percent of GDP in 2086.*

The Social Security Shortfall Social Security is the major public retirement program in the United States. The program costs $736 billion annually and is financed by a 12.4 percent tax on earnings up to a set level of earnings ($113,700 in 2013). Half the tax (6.2 percent) is paid by the worker; the other half by the employer. Social Security is largely an annual "pay-as-you-go" plan, meaning that most of the current revenues from the Social Security tax are paid to current Social Security retirees. Through the start of 2009, Social Security revenues exceeded Social Security payouts in anticipation of the large benefits promised to the baby boomers when they retire. That excess inflow was used to buy U.S. Treasury securities that were credited to a government account called the Social Security Trust Fund. But the combined accumulation of money in the trust fund through 2009 plus the projected future revenues from the payroll tax in later years was expected to be greatly inadequate for paying the promised retirement benefits to all future retirees.

This underfunding of future retirement promises was brought into sharper focus in the latter half of 2009 when, for the first time, Social Security revenues fell below Social Security retirement payouts and the system started shifting money from the trust fund to make up the difference. The trust fund is expected to be exhausted in 2033. For each year thereafter, annual tax revenues will cover only 75 percent of the promised benefits.

*Social Security and Medicare Board of Trustees, "Status of the Social Security and Medicare Programs: A Summary of the 2012 Annual Reports," **www.ssa.gov**. This publication is also the source of most of the statistical information that follows.

SUMMARY

LO31.1 Identify and explain the purposes, tools, and limitations of fiscal policy.

Fiscal policy consists of deliberate changes in government spending, taxes, or some combination of both to promote full employment, price-level stability, and economic growth. Fiscal policy requires increases in government spending, decreases in taxes, or both—a budget deficit—to increase aggregate demand and push an economy from a recession. Decreases in government spending, increases in taxes, or both—a budget surplus—are appropriate fiscal policy for decreasing aggregate demand to try to slow or halt demand-pull inflation.

The Medicare Shortfall The Medicare program is the U.S. health care program for people age 65 and older in the United States. The program costs $558 billion per year and has been growing at about 9 percent annually. Like Social Security, it also is a pay-as-you-go plan, meaning that current medical benefits for people age 65 or older are being funded by current tax revenues from the 2.9 percent Medicare tax on earnings and the 3.8 percent investment surtax on wealthier investors that was instituted in 2013 as part of Obamacare. Like the Social Security tax, half the Medicare earnings tax is paid by the employer (1.45 percent) and half the by the employee. But the 2.9 percent Medicare tax is applied to all earnings.

The financial status of Medicare is much worse than that of Social Security. To begin with, the Medicare Trust Fund will be depleted in 2024, 9 years before the Social Security Trust Fund is expected to be depleted. Then, in subsequent years, the percentage of scheduled Medicare benefits covered by the Medicare tax will decline from 97 percent in 2025 to 72 percent in 2035 and 69 percent in 2080.

The Unpleasant Options To restore long-run balance to Social Security and Medicare, the federal government must either reduce benefits or increase revenues. It's as simple—and as complicated—as that! The Social Security Administration concludes that bringing projected Social Security revenues and payments into balance over the next 75 years would require a 16 percent permanent reduction in Social Security benefits, a 13 percent permanent increase in tax revenues, or some combination of the two. To bring projected Medicare revenues and expenses into long-run balance would require an increase in the Medicare payroll tax by 122 percent, a 51 percent reduction of Medicare payments from their projected levels, or some combination of each.

All the general options for closing all or part of the Social Security and Medicare gaps involve difficult economic trade-offs and dangerous political risks because one group or another will strongly oppose them. Here are just a few examples:

- Increasing the retirement age for collecting Social Security or Medicare benefits will upset preretirement workers who have been paying into the system and will receive their benefits later than they expected.
- Subjecting a larger portion of total earnings to the Social Security tax would constitute a gigantic tax increase on the earnings of the country's highest trained and educated individuals. This might reduce the incentive of younger people to obtain education and advance in their careers.
- Disqualifying wealthy individuals from receiving Social Security and Medicare benefits would tilt the programs toward welfare and redistribution, rather than insurance programs. This would undermine the broad existing political support for the programs.
- Redirecting legal immigration toward high-skilled, high-earning entrants and away from low-skilled, low-earning immigrants to raise Social Security and Medicare revenues would also raise the ire of some native-born high-skilled workers and the proponents of immigration based on family reunification.
- Placing the payroll tax revenues into accounts that individuals, not the government, would own, maintain, and bequeath would transform the Social Security and Medicare programs from guaranteed "defined benefit plans" into much riskier "defined contribution plans." The extreme short-run volatility of the stock market might leave some unlucky people destitute in old age.

The problem is huge and will not go away. One recent attempt to add up the underfunding of all the promised Social Security and Medicare benefits finds that the total underfunding greatly exceeds the combined current wealth (net worth) of everyone in the United States today.[†] Even if this estimate is somewhat extreme, the fact remains: The federal government and American people eventually will have to face up to the over-promising-underfunding problem and find ways to resolve it.

[†]Bruce Bartlett, "The 81% Tax Increase," Forbes.com, August 8, 2009, **www.forbes.com**.

LO31.2 Explain the role of built-in stabilizers in moderating business cycles.

Built-in stability arises from net tax revenues, which vary directly with the level of GDP. During recession, the federal budget automatically moves toward a stabilizing deficit; during expansion, the budget automatically moves toward an anti-inflationary surplus. Built-in stability lessens, but does not fully correct, undesired changes in real GDP.

LO31.3 Describe how the cyclically adjusted budget reveals the status of U.S. fiscal policy.

Actual federal budget deficits can go up or down because of changes in GDP, changes in fiscal policy, or both. Deficits caused by changes in GDP are called cyclical deficits. The cyclically adjusted budget removes cyclical deficits from the budget and therefore measures the budget deficit or surplus that would occur if the economy operated at its full-employment output throughout the year. Changes in the cyclical-budget deficit or surplus provide meaningful information as to whether the government's fiscal policy is expansionary, neutral, or contractionary. Changes in the actual budget deficit or surplus do not, since such deficits or surpluses can include cyclical deficits or surpluses.

LO31.4 Summarize recent U.S. fiscal policy and the projections for U.S. fiscal policy over the next few years.

In 2001 the Bush administration and Congress chose to reduce marginal tax rates and phase out the federal estate tax. A recession occurred in 2001, and federal spending for the war on terrorism rocketed. The federal budget swung from a surplus of $128 billion in 2001 to a deficit of $158 billion in 2002. In 2003 the Bush administration and Congress accelerated the tax reductions scheduled under the 2001 tax law and cut tax rates on capital gains and dividends. The purposes were to stimulate a sluggish economy. By 2007 the economy had reached its full employment level of output.

The federal government responded to the deep recession of 2007–2009 by implementing highly expansionary fiscal policy. In 2008 the federal government passed a tax rebate program that sent $600 checks to qualified individuals. Later that year, it created a $700 billion emergency fund to keep key financial institutions from failing. These and other programs increased the cyclically adjusted budget deficit from −1.3 percent of potential GDP in 2007 to −2.9 percent in 2008. When the economy continued to plunge, the Obama administration and Congress enacted a massive $787 billion stimulus program to be implemented over 2½ years. The cyclically adjusted budget deficit shot up from −2.9 percent of potential GDP in 2008 to −7.1 percent in 2009.

LO31.5 Discuss the problems that governments may encounter in enacting and applying fiscal policy.

Certain problems complicate the enactment and implementation of fiscal policy. They include (*a*) timing problems associated with recognition, administrative, and operational lags; (*b*) the potential for misuse of fiscal policy for political rather than economic purposes; (*c*) the fact that state and local finances tend to be pro-cyclical; (*d*) potential ineffectiveness if households expect future policy reversals; and (*e*) the possibility of fiscal policy crowding out private investment.

Most economists believe that fiscal policy can help move the economy in a desired direction but cannot reliably be used to fine-tune the economy to a position of price stability and full employment. Nevertheless, fiscal policy is a valuable backup tool for aiding monetary policy in fighting significant recession or inflation.

LO31.6 Discuss the size, composition, and consequences of the U.S. public debt.

The public debt is the total accumulation of all past federal government deficits and surpluses and consists of Treasury bills, Treasury notes, Treasury bonds, and U.S. savings bonds. In 2012 the U.S. public debt was $16.4 trillion, or $52,396 per person. The public (which here includes banks and state and local governments) holds 60 percent of that federal debt; the Federal Reserve and federal agencies hold the other 40 percent. Foreigners hold 33 percent of the federal debt. Interest payments as a percentage of GDP were about 2.3 percent in 2012. Because of large deficits during the Great Recession and in subsequent years, the total U.S. public debt passed $17.0 trillion in 2013. It had doubled in seven years.

The concern that a large public debt may bankrupt the U.S. government is generally a false worry because (*a*) the debt needs only to be refinanced rather than refunded and (*b*) the federal government has the power to increase taxes to make interest payments on the debt.

In general, the public debt is not a vehicle for shifting economic burdens to future generations. Americans inherit not only most of the public debt (a liability) but also most of the U.S. government securities (an asset) that finance the debt.

More substantive problems associated with public debt include the following: (*a*) Payment of interest on the debt may increase income inequality. (*b*) Interest payments on the debt require higher taxes, which may impair incentives. (*c*) Paying interest or principal on the portion of the debt held by foreigners means a transfer of real output abroad. (*d*) Government borrowing to refinance or pay interest on the debt may increase interest rates and crowd out private investment spending, leaving future generations with a smaller stock of capital than they would have had otherwise.

The increase in investment in public capital that may result from debt financing may partly or wholly offset the crowding-out effect of the public debt on private investment. Also, the added public investment may stimulate private investment, where the two are complements.

TERMS AND CONCEPTS

fiscal policy
Council of Economic Advisers (CEA)
expansionary fiscal policy
budget deficit
contractionary fiscal policy
budget surplus
built-in stabilizer
progressive tax system
proportional tax system
regressive tax system
cyclically adjusted budget
cyclical deficit
political business cycle
crowding-out effect
public debt
U.S. government securities
external public debt
public investments

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What is the role of the Council of Economic Advisers (CEA) as it relates to fiscal policy? Use an Internet search to find the names and university affiliations of the present members of the CEA. **LO31.1**
2. What are government's fiscal policy options for ending severe demand-pull inflation? Which of these fiscal options do you think might be favored by a person who wants to preserve the size of government? A person who thinks the public sector is too large? How does the "ratchet effect" affect anti-inflationary fiscal policy? **LO31.1**
3. (For students who were assigned Chapter 29) Use the aggregate expenditures model to show how government fiscal policy could eliminate either a recessionary expenditure gap or an inflationary expenditure gap (Figure 29.7). Explain how equal-size increases in G and T could eliminate a recessionary gap and how equal-size decreases in G and T could eliminate an inflationary gap. **LO31.1**
4. Some politicians have suggested that the United States enact a constitutional amendment requiring that the federal government balance its budget annually. Explain why such an amendment, if strictly enforced, would force the government to enact a contractionary fiscal policy whenever the economy experienced a severe recession. **LO31.1**
5. Explain how built-in (or automatic) stabilizers work. What are the differences between proportional, progressive, and regressive tax systems as they relate to an economy's built-in stability? **LO31.2**
6. Define the cyclically adjusted budget, explain its significance, and state why it may differ from the actual budget. Suppose the full-employment, noninflationary level of real output is GDP_3 (not GDP_2) in the economy depicted in Figure 31.3. If the economy is operating at GDP_2, instead of GDP_3, what is the status of its cyclically adjusted budget? The status of its current fiscal policy? What change in fiscal policy would you recommend? How would you accomplish that in terms of the G and T lines in the figure? **LO31.3**
7. Briefly state and evaluate the problem of time lags in enacting and applying fiscal policy. Explain the idea of a political business cycle. How might expectations of a near-term policy reversal weaken fiscal policy based on changes in tax rates? What is the crowding-out effect, and why might it be relevant to fiscal policy? In view of your answers, explain the following statement: "Although fiscal policy clearly is useful in combating the extremes of severe recession and demand-pull inflation, it is impossible to use fiscal policy to fine-tune the economy to the full-employment, noninflationary level of real GDP and keep the economy there indefinitely." **LO31.5**
8. How do economists distinguish between the absolute and relative sizes of the public debt? Why is the distinction important? Distinguish between refinancing the debt and retiring the debt. How does an internally held public debt differ from an externally held public debt? Contrast the effects of retiring an internally held debt and retiring an externally held debt. **LO31.6**
9. True or false? If false, explain why. **LO31.6**
   a. The total public debt is more relevant to an economy than the public debt as a percentage of GDP.
   b. An internally held public debt is like a debt of the left hand owed to the right hand.
   c. The Federal Reserve and federal government agencies hold more than three-fourths of the public debt.
   d. The portion of the U.S. debt held by the public (and not by government entities) was larger as a percentage of GDP in 2012 than it was in 2000.
   e. As a percentage of GDP, the total U.S. public debt is the highest such debt among the world's advanced industrial nations.
10. Why might economists be quite concerned if the annual interest payments on the U.S. public debt sharply increased as a percentage of GDP? **LO31.6**
11. Trace the cause-and-effect chain through which financing and refinancing of the public debt might affect real interest rates, private investment, the stock of capital, and economic growth. How might investment in public capital and complementarities between public capital and private capital alter the outcome of the cause-effect chain? **LO31.6**
12. **LAST WORD** What do economists mean when they say Social Security and Medicare are "pay-as-you-go" plans? What are the Social Security and Medicare trust funds, and how long will they have money left in them? What is the key long-run problem of both Social Security and Medicare? Do you favor increasing taxes or do you prefer reducing benefits to fix the problem?

REVIEW QUESTIONS

1. Which of the following would help a government reduce an inflationary output gap? **LO31.1**
   a. Raising taxes.
   b. Lowering taxes.
   c. Increasing government spending.
   d. Decreasing government spending.
2. The economy is in a recession. A congresswoman suggests increasing spending to stimulate aggregate demand but also at the same time raising taxes to pay for the increased spending. Her suggestion to combine higher government expenditures with higher taxes is: **LO31.1**
   a. The worst possible combination of tax and expenditure changes.
   b. The best possible combination of tax and expenditure changes.

   c. A mediocre and contradictory combination of tax and expenditure changes.
   d. None of the above.

3. During the recession of 2007–2009, the U.S. federal government's tax collections fell from about $2.6 trillion down to about $2.1 trillion while GDP declined by about 4 percent. Does the U.S. tax system appear to have built-in stabilizers? **LO31.2**
   a. Yes.
   b. No.
4. Last year, while an economy was in a recession, government spending was $595 billion and government revenue was $505 billion. Economists estimate that if the economy had been at its full-employment level of GDP last year, government spending would have been $555 billion and government revenue would have been $550 billion. Which of the following statements about this government's fiscal situation are true? **LO31.3**
   a. The government has a non–cyclically adjusted budget deficit of $595 billion.
   b. The government has a non–cyclically adjusted budget deficit of $90 billion.
   c. The government has a non–cyclically adjusted budget surplus of $90 billion.
   d. The government has a cyclically adjusted budget deficit of $555 billion.
   e. The government has a cyclically adjusted budget deficit of $5 billion.
   f. The government has a cyclically adjusted budget surplus of $5 billion.
5. Label each of the following scenarios in which there are problems enacting and applying fiscal policy as being an example of either recognition lag, administrative lag, or operational lag. **LO31.5**
   a. To fight a recession, Congress has passed a bill to increase infrastructure spending—but the legally required environmental-impact statement for each new project will take at least two years to complete before any building can begin.
   b. Distracted by a war that is going badly, inflation reaches 8 percent before politicians take notice.
   c. A sudden recession is recognized by politicians, but it takes many months of political deal making before a stimulus bill is finally approved.
   d. To fight a recession, the president orders federal agencies to get rid of petty regulations that burden private businesses—but the federal agencies begin by spending a year developing a set of regulations on how to remove petty regulations.
6. In January, the interest rate is 5 percent and firms borrow $50 billion per month for investment projects. In February, the federal government doubles its monthly borrowing from $25 billion to $50 billion. That drives the interest rate up to 7 percent. As a result, firms cut back their borrowing to only $30 billion per month. Which of the following is true? **LO31.6**
   a. There is no crowding-out effect because the government's increase in borrowing exceeds firm's decrease in borrowing.
   b. There is a crowding-out effect of $20 billion.
   c. There is no crowding-out effect because both the government and firms are still borrowing a lot.
   d. There is a crowding-out effect of $25 billion.

PROBLEMS

1. Assume that a hypothetical economy with an MPC of .8 is experiencing severe recession. By how much would government spending have to rise to shift the aggregate demand curve rightward by $25 billion? How large a tax cut would be needed to achieve the same increase in aggregate demand? Determine one possible combination of government spending increases and tax decreases that would accomplish the same goal. **LO31.1**
2. Refer back to the table in Figure 30.7 in the previous chapter. Suppose that aggregate demand increases such that the amount of real output demanded rises by $7 billion at each price level. By what percentage will the price level increase? Will this inflation be demand-pull inflation or will it be cost-push inflation? If potential real GDP (that is, full-employment GDP) is $510 billion, what will be the size of the positive GDP gap after the change in aggregate demand? If government wants to use fiscal policy to counter the resulting inflation without changing tax rates, would it increase government spending or decrease it? **LO31.1**
3. (For students who were assigned Chapter 29) Assume that, without taxes, the consumption schedule for an economy is as shown below: **LO31.1**

GDP, Billions	Consumption, Billions
$100	$120
200	200
300	280
400	360
500	440
600	520
700	600

   a. Graph this consumption schedule. What is the size of the MPC?
   b. Assume that a lump-sum (regressive) tax of $10 billion is imposed at all levels of GDP. Calculate the tax rate at

each level of GDP. Graph the resulting consumption schedule and compare the MPC and the multiplier with those of the pretax consumption schedule.

c. Now suppose a proportional tax with a 10 percent tax rate is imposed instead of the regressive tax. Calculate and graph the new consumption schedule and note the MPC and the multiplier.

d. Finally, impose a progressive tax such that the tax rate is 0 percent when GDP is $100, 5 percent at $200, 10 percent at $300, 15 percent at $400, and so forth. Determine and graph the new consumption schedule, noting the effect of this tax system on the MPC and the multiplier.

e. Use a graph similar to Figure 31.3 to show why proportional and progressive taxes contribute to greater economic stability, while a regressive tax does not.

4. Refer to the following table for Waxwania: **LO31.2**

Government Expenditures, *G*	Tax Revenues, *T*	Real GDP
$160	$100	$500
160	120	600
160	140	700
160	160	800
160	180	900

What is the marginal tax rate in Waxwania? The average tax rate? Which of the following describes the tax system: proportional, progressive, regressive?

5. Refer to the table for Waxwania in problem 4. Suppose that Waxwania is producing $600 of real GDP, whereas the potential real GDP (or full-employment real GDP) is $700. How large is its budget deficit? Its cyclically adjusted budget deficit? Its cyclically adjusted budget deficit as a percentage of potential real GDP? Is Waxwania's fiscal policy expansionary or is it contractionary? **LO31.3**

6. Suppose that a country has no public debt in year 1 but experiences a budget deficit of $40 billion in year 2, a budget surplus of $10 billion in year 3, and a budget deficit of $2 billion in year 4. What is the absolute size of its public debt in year 4? If its real GDP in year 4 is $104 billion, what is this country's public debt as a percentage of real GDP in year 4? **LO31.6**

7. Suppose that the investment demand curve in a certain economy is such that investment declines by $100 billion for every 1 percentage point increase in the real interest rate. Also, suppose that the investment demand curve shifts rightward by $150 billion at each real interest rate for every 1 percentage point increase in the expected rate of return from investment. If stimulus spending (an expansionary fiscal policy) by government increases the real interest rate by 2 percentage points, but also raises the expected rate of return on investment by 1 percentage point, how much investment, if any, will be crowded out? **LO31.6**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **NINE**

MONEY, BANKING, AND MONETARY POLICY

CHAPTER **32**

Money, Banking, and Financial Institutions

Learning Objectives

LO32.1 Identify and explain the functions of money.

LO32.2 List and describe the components of the U.S. money supply.

LO32.3 Describe what "backs" the money supply, making us willing to accept it as payment.

LO32.4 Discuss the makeup of the Federal Reserve and its relationship to banks and thrifts.

LO32.5 Identify the functions and responsibilities of the Federal Reserve and explain why Fed independence is important.

LO32.6 Identify and explain the main factors that contributed to the financial crisis of 2007–2008.

LO32.7 Discuss the actions of the U.S. Treasury and the Federal Reserve that helped keep the banking and financial crisis of 2007–2008 from worsening.

LO32.8 Identify the main subsets of the financial services industry in the United States and provide examples of some firms in each category.

Money is a fascinating aspect of the economy:

> Money bewitches people. They fret for it, and they sweat for it. They devise most ingenious ways to get it, and most ingenuous ways to get rid of it. Money is the only commodity that is good for nothing but to be gotten rid of. It will not feed you, clothe you, shelter you, or amuse you unless you spend it or invest it. It

imparts value only in parting. People will do almost anything for money, and money will do almost anything for people. Money is a captivating, circulating, masquerading puzzle.[1]

In this chapter and the two chapters that follow, we want to unmask the critical role of money and the monetary system in the economy. When the monetary system is working properly, it provides the lifeblood of the circular flows of income and expenditure. A well-operating monetary system helps the economy achieve both full employment and the efficient use of resources. A malfunctioning monetary system distorts the allocation of resources and creates severe fluctuations in the economy's levels of output, employment, and prices.

[1]"Creeping Inflation," *Business Review,* August 1957, p. 3. Federal Reserve Bank of Philadelphia. Used with permission.

The Functions of Money

LO32.1 Identify and explain the functions of money.

Just what is money? There is an old saying that "money *is* what money *does*." In a general sense, anything that performs the functions of money *is* money. Here are those functions:

- ***Medium of exchange*** First and foremost, money is a **medium of exchange** that is usable for buying and selling goods and services. A bakery worker does not want to be paid 200 bagels per week. Nor does the bakery owner want to receive, say, halibut in exchange for bagels. Money, however, is readily acceptable as payment. As we saw in Chapter 2, money is a social invention with which resource suppliers and producers can be paid and that can be used to buy any of the full range of items available in the marketplace. As a medium of exchange, money allows society to escape the complications of barter. And because it provides a convenient way of exchanging goods, money enables society to gain the advantages of geographic and human specialization.
- ***Unit of account*** Money is also a **unit of account.** Society uses monetary units—dollars, in the United States—as a yardstick for measuring the relative worth of a wide variety of goods, services, and resources. Just as we measure distance in miles or kilometers, we gauge the value of goods in dollars.

  With money as an acceptable unit of account, the price of each item need be stated only in terms of the monetary unit. We need not state the price of cows in terms of corn, crayons, and cranberries. Money aids rational decision making by enabling buyers and sellers to easily compare the prices of various goods, services, and resources. It also permits us to define debt obligations, determine taxes owed, and calculate the nation's GDP.
- ***Store of value*** Money also serves as a **store of value** that enables people to transfer purchasing power from the present to the future. People normally do not spend all their incomes on the day they receive them. To buy things later, they store some of their wealth as money. The money you place in a safe or a checking account will still be available to you a few weeks or months from now. When inflation is nonexistent or mild, holding money is a relatively risk-free way to store your wealth for later use.

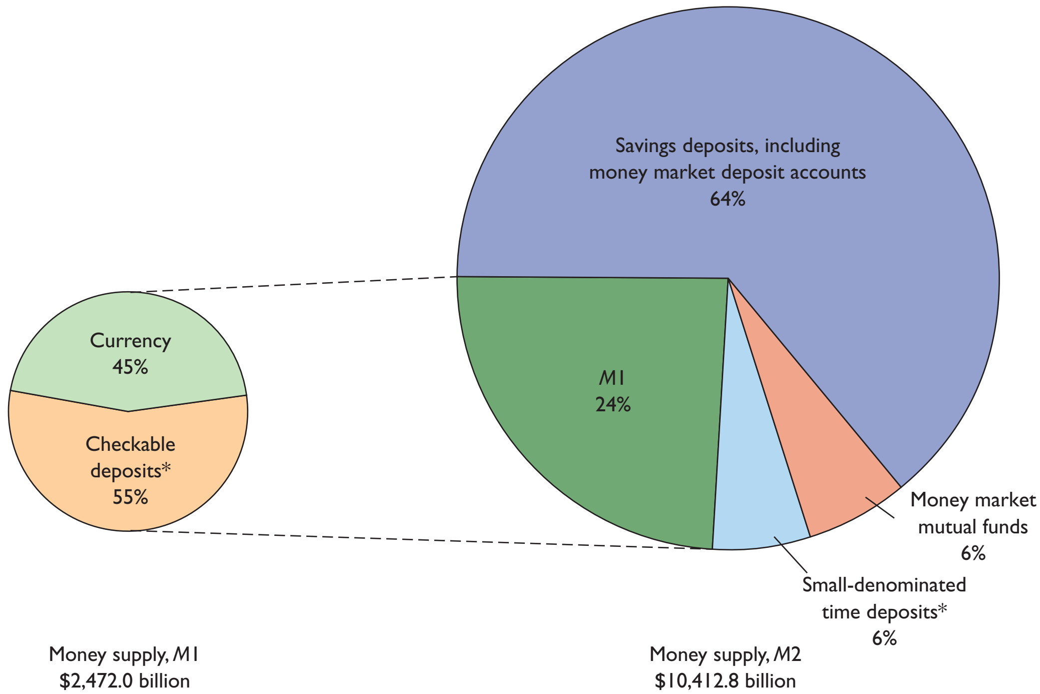
People can, of course, choose to hold some or all of their wealth in a wide variety of assets besides money. These include real estate, stocks, bonds, precious metals such as gold, and even collectible items like fine art or comic books. But a key advantage that money has over all other assets is that it has the most *liquidity*, or spendability.

An asset's **liquidity** is the ease with which it can be converted quickly into the most widely accepted and easily spent form of money, cash, with little or no loss of purchasing power. The more liquid an asset is, the more quickly it can be converted into cash and used for either purchases of goods and services or purchases of other assets.

Levels of liquidity vary radically. By definition, cash is perfectly liquid. By contrast, a house is highly illiquid for two reasons. First, it may take several months before a willing buyer can be found and a sale negotiated so that its value can be converted into cash. Second, there is a loss of purchasing power when the house is sold because numerous fees have to be paid to real estate agents and other individuals to complete the sale.

As we are about to discuss, our economy uses several different types of money including cash, coins, checking

FIGURE 32.1 Components of money supply $M1$ and money supply $M2$, in the United States. (a) $M1$ is a narrow definition of the money supply that includes currency (in circulation) and checkable deposits. (b) $M2$ is a broader definition that includes $M1$ along with several other relatively liquid account balances.



*These categories include other, quantitatively smaller components such as traveler's checks.

Source: Federal Reserve System, **www.federalreserve.gov**. Data are for February 2013.

account deposits, savings account deposits, and even more exotic things like deposits in money market mutual funds. As we describe the various forms of money in detail, take the time to compare their relative levels of liquidity—both with each other and as compared to other assets like stocks, bonds, and real estate. Cash is perfectly liquid. Other forms of money are highly liquid, but less liquid than cash.

The Components of the Money Supply

LO32.2 List and describe the components of the U.S. money supply.

Money is a "stock" of some item or group of items (unlike income, for example, which is a "flow"). Societies have used many items as money, including whales' teeth, circular stones, elephant-tail bristles, gold coins, furs, and pieces of paper. Anything that is widely accepted as a medium of exchange can serve as money. In the United States, currency is not the only form of money. As you will see, certain debts of government and financial institutions also are used as money.

Money Definition *M*1

The narrowest definition of the U.S. money supply is called ***M*1.** It consists of two components:

- Currency (coins and paper money) in the hands of the public.
- All checkable deposits (all deposits in commercial banks and "thrift" or savings institutions on which checks of any size can be drawn).[2]

Government and government agencies supply coins and paper money. Commercial banks ("banks") and savings institutions ("thrifts") provide checkable deposits. Figure 32.1a shows that $M1$ is about equally divided between the two components.

[2]In the ensuing discussion, we do not discuss several of the quantitatively less significant components of the definitions of money to avoid a maze of details. For example, traveler's checks are included in the $M1$ money supply. The statistical appendix of any recent *Federal Reserve Bulletin* provides more comprehensive definitions.

Currency: Coins + Paper Money The currency of the United States consists of metal coins and paper money. The coins are issued by the U.S. Treasury while the paper money consists of **Federal Reserve Notes** issued by the Federal Reserve System (the U.S. central bank). The coins are minted by the U.S. Mint while the paper money is printed by the Bureau of Engraving and Printing. Both the U.S. Mint and the Bureau of Engraving and Printing are part of the U.S. Department of the Treasury.

As with the currencies of other countries, the currency of the United States is **token money.** This means that the face value of any piece of currency is unrelated to its *intrinsic value*—the value of the physical material (metal or paper and ink) out of which that piece of currency is constructed. Governments make sure that face values exceed intrinsic values to discourage people from destroying coins and bills to resell the material that they are made out of. For instance, if 50-cent pieces each contained 75 cents' worth of metal, then it would be profitable to melt them down and sell the metal. Fifty-cent pieces would disappear from circulation very quickly!

Figure 32.1a shows that currency (coins and paper money) constitutes 45 percent of the $M1$ money supply in the United States.

Checkable Deposits The safety and convenience of checks has made **checkable deposits** a large component of the $M1$ money supply. You would not think of stuffing $4,896 in bills in an envelope and dropping it in a mailbox to pay a debt. But writing and mailing a check for a large sum is commonplace. The person cashing a check must endorse it (sign it on the reverse side); the writer of the check subsequently receives a record of the cashed check as a receipt attesting to the fulfillment of the obligation. Similarly, because the writing of a check requires endorsement, the theft or loss of your checkbook is not nearly as calamitous as losing an identical amount of currency. Finally, it is more convenient to write a check than to transport and count out a large sum of currency. For all these reasons, checkable deposits (checkbook money) are a large component of the stock of money in the United States. About 55 percent of $M1$ is in the form of checkable deposits, on which checks can be drawn.

It might seem strange that checking account balances are regarded as part of the money supply. But the reason is clear: Checks are nothing more than a way to transfer the ownership of deposits in banks and other financial institutions and are generally acceptable as a medium of exchange. Although checks are less generally accepted than currency for small purchases, for major purchases most sellers willingly accept checks as payment. Moreover, people can convert checkable deposits into paper money and coins on demand; checks drawn on those deposits are thus the equivalent of currency.

To summarize:

$$\text{Money, } M1 = \text{currency} + \text{checkable deposits}$$

Institutions That Offer Checkable Deposits In the United States, a variety of financial institutions allow customers to write checks in any amount on the funds they have deposited. **Commercial banks** are the primary depository institutions. They accept the deposits of households and businesses, keep the money safe until it is demanded via checks, and in the meantime use it to make available a wide variety of loans. Commercial bank loans provide short-term financial capital to businesses, and they finance consumer purchases of automobiles and other durable goods.

Savings and loan associations (S&Ls), mutual savings banks, and credit unions supplement the commercial banks and are known collectively as savings or **thrift institutions,** or simply "thrifts." *Savings and loan associations* and *mutual savings banks* accept the deposits of households and businesses and then use the funds to finance housing mortgages and to provide other loans. *Credit unions* accept deposits from and lend to "members," who usually are a group of people who work for the same company.

The checkable deposits of banks and thrifts are known variously as demand deposits, NOW (negotiable order of withdrawal) accounts, ATS (automatic transfer service) accounts, and share draft accounts. Their commonality is that depositors can write checks on them whenever, and in whatever amount, they choose.

Two Qualifications We must qualify our discussion in two important ways. First, currency held by the U.S. Treasury, the Federal Reserve banks, commercial banks, and thrift institutions is *excluded* from $M1$ and other measures of the money supply. A paper dollar or four quarters in the billfold of, say, Emma Buck obviously constitutes just $1 of the money supply. But if we counted currency held by banks as part of the money supply, the same $1 would count for $2 of money supply when Emma deposited the currency into her checkable deposit in her bank. It would count for $1 of checkable deposit owned by Buck and also $1 of currency in the bank's cash drawer or vault. By excluding currency held by banks when determining the total supply of money, we avoid this problem of double counting.

Also *excluded* from the money supply are any checkable deposits of the government (specifically, the U.S. Treasury) or the Federal Reserve that are held by commercial banks or thrift institutions. This exclusion is designed to enable a better assessment of the amount of money available *to the private sector* for potential spending. The amount of money available to households and businesses is of keen interest to the Federal Reserve in conducting its monetary policy (a topic we cover in detail in Chapter 34).

Money Definition *M*2

A second and broader definition of money includes *M*1 plus several near-monies. **Near-monies** are certain highly liquid financial assets that do not function directly or fully as a medium of exchange but can be readily converted into currency or checkable deposits. The ***M*2** definition of money includes three categories of near-monies.

- ***Savings deposits, including money market deposit accounts*** A depositor can easily withdraw funds from a **savings account** at a bank or thrift or simply request that the funds be transferred from a savings account to a checkable account. A person can also withdraw funds from a **money market deposit account (MMDA),** which is an interest-bearing account containing a variety of interest-bearing short-term securities. MMDAs, however, have a minimum-balance requirement and a limit on how often a person can withdraw funds.
- ***Small-denominated (less than $100,000) time deposits*** Funds from **time deposits** become available at their maturity. For example, a person can convert a 6-month time deposit ("certificate of deposit," or "CD") to currency without penalty 6 months or more after it has been deposited. In return for this withdrawal limitation, the financial institution pays a higher interest rate on such deposits than it does on its MMDAs. Also, a person can "cash in" a CD at any time but must pay a severe penalty.
- ***Money market mutual funds held by individuals*** By making a telephone call, using the Internet, or writing a check for $500 or more, a depositor can redeem shares in a **money market mutual fund (MMMF)** offered by a mutual fund company. Such companies use the combined funds of individual shareholders to buy interest-bearing short-term credit instruments such as certificates of deposit and U.S. government securities. Then they can offer interest on the MMMF accounts of the shareholders (depositors) who jointly own those financial assets. The MMMFs in *M*2 include only the MMMF accounts held by individuals; those held by businesses and other institutions are excluded.

All three categories of near-monies imply substantial liquidity. Thus, in equation form,

$$\text{Money, } M2 = \begin{array}{l} M1 + \text{savings deposits, including} \\ \text{MMDAs} + \text{small-denominated} \\ \text{(less than \$100,000) time deposits} \\ + \text{MMMFs held by individuals} \end{array}$$

In summary, *M*2 includes the immediate medium-of-exchange items (currency and checkable deposits) that constitute *M*1 plus certain near-monies that can be easily

CONSIDER THIS . . .

Are Credit Cards Money?

You may wonder why we have ignored credit cards such as Visa and MasterCard in our discussion of how the money supply is defined. After all, credit cards are a convenient way to buy things and account for about 25 percent of the dollar value of all transactions in the United States. The answer is that a credit card is not money. Rather, it is a convenient means of obtaining a short-term loan from the financial institution that issued the card.

What happens when you purchase an item with a credit card? The bank that issued the card will reimburse the seller by making a money payment and charging the establishment a transaction fee, and later you will reimburse the bank for its loan to you by also making a money payment. Rather than reduce your cash or checking account with each purchase, you bunch your payments once a month. You may have to pay an annual fee for the services provided, and if you pay the bank in installments, you will pay a sizable interest charge on the loan. Credit cards are merely a means of deferring or postponing payment for a short period. Your checking account balance that you use to pay your credit card bill *is* money; the credit card is *not* money.*

Although credit cards are not money, they allow individuals and businesses to "economize" in the use of money. Credit cards enable people to hold less currency in their billfolds and, prior to payment due dates, fewer checkable deposits in their bank accounts. Credit cards also help people coordinate the timing of their expenditures with their receipt of income.

*A bank debit card, however, is very similar to a check in your checkbook. Unlike a purchase with a credit card, a purchase with a debit card creates a direct "debit" (a subtraction) from your checking account balance. That checking account balance is money—it is part of *M*1.

converted into currency and checkable deposits. In Figure 32.1b we see that the addition of all these items yields an *M*2 money supply that is about five times larger than the narrower *M*1 money supply.

QUICK REVIEW 32.1

- Money serves as a medium of exchange, a unit of account, and a store of value.
- The narrow *M*1 definition of money includes currency held by the public plus checkable deposits in commercial banks and thrift institutions.
- Thrift institutions as well as commercial banks offer accounts on which checks can be written.
- The *M*2 definition of money includes *M*1 plus savings deposits, including money market deposit accounts, small-denominated (less than $100,000) time deposits, and money market mutual fund balances held by individuals.

What "Backs" the Money Supply?

LO32.3 Describe what "backs" the money supply, making us willing to accept it as payment.

The money supply in the United States essentially is "backed" (guaranteed) by the government's ability to keep the value of money relatively stable. Nothing more!

Money as Debt

The major components of the money supply—paper money and checkable deposits—are debts, or promises to pay. In the United States, paper money is the circulating debt of the Federal Reserve Banks. Checkable deposits are the debts of commercial banks and thrift institutions.

Paper currency and checkable deposits have no intrinsic value. A $5 bill is just an inscribed piece of paper. A checkable deposit is merely a bookkeeping entry. And coins, we know, have less intrinsic value than their face value. Nor will government redeem the paper money you hold for anything tangible, such as gold. To many people, the fact that the government does not back the currency with anything tangible seems implausible and insecure. But the decision not to back the currency with anything tangible was made for a very good reason. If the government backed the currency with something tangible like gold, then the supply of money would vary with how much gold was available. By not backing the currency, the government avoids this constraint and indeed receives a key freedom—the ability to provide as much or as little money as needed to maintain the value of money and to best suit the economic needs of the country. In effect, by choosing not to back the currency, the government has chosen to give itself the ability to freely "manage" the nation's money supply. Its monetary authorities attempt to provide the amount of money needed for the particular volume of business activity that will promote full employment, price-level stability, and economic growth.

Nearly all today's economists agree that managing the money supply is more sensible than linking it to gold or to some other commodity whose supply might change arbitrarily and capriciously. For instance, if we used gold to back the money supply so that gold was redeemable for money and vice versa, then a large increase in the nation's gold stock as the result of a new gold discovery might increase the money supply too rapidly and thereby trigger rapid inflation. Or a long-lasting decline in gold production might reduce the money supply to the point where recession and unemployment resulted.

In short, people cannot convert paper money into a fixed amount of gold or any other precious commodity. Money is exchangeable only for paper money. If you ask the government to redeem $5 of your paper money, it will swap one paper $5 bill for another bearing a different serial number. That is all you can get. Similarly, checkable deposits can be redeemed not for gold but only for paper money, which, as we have just seen, the government will not redeem for anything tangible.

Value of Money

So why are currency and checkable deposits money, whereas, say, Monopoly (the game) money is not? What gives a $20 bill or a $100 checking account entry its value? The answer to these questions has three parts.

Acceptability Currency and checkable deposits are money because people accept them as money. By virtue of long-standing business practice, currency and checkable deposits perform the basic function of money: They are acceptable as a medium of exchange. We accept paper money in exchange because we are confident it will be exchangeable for real goods, services, and resources when we spend it.

Legal Tender Our confidence in the acceptability of paper money is strengthened because the government has designated currency as **legal tender.** Specifically, each bill contains the statement "This note is legal tender for all debts, public and private." That means paper money is a valid and legal means of payment of any debt that was contracted in dollars. (But private firms and government are not mandated to accept cash. It is not illegal for them to specify payment in noncash forms such as checks, cashier's checks, money orders, or credit cards.)

The general acceptance of paper currency in exchange is more important than the government's decree that money is legal tender, however. The government has never decreed checks to be legal tender, and yet they serve as such in many of the economy's exchanges of goods, services, and resources. But it is true that government agencies—the Federal Deposit Insurance Corporation (FDIC) and the National Credit Union Administration (NCUA)—insure individual deposits of up to $250,000 at commercial banks and thrifts. That fact enhances our willingness to use checkable deposits as a medium of exchange.

Relative Scarcity The value of money, like the economic value of anything else, depends on its supply and demand. Money derives its value from its scarcity relative to its utility (its want-satisfying power). The utility of money lies in its capacity to be exchanged for goods and services, now or in the future. The economy's demand for money thus depends on the total dollar volume of transactions in any period plus the amount of money individuals and businesses want to hold for future transactions. With a reasonably constant demand for money, the supply of money provided by the monetary authorities will determine the domestic value or "purchasing power" of the monetary unit (dollar, yen, peso, or whatever).

Money and Prices

The purchasing power of money is the amount of goods and services a unit of money will buy. When money rapidly loses its purchasing power, it loses its role as money.

The Purchasing Power of the Dollar The amount a dollar will buy varies inversely with the price level; that is, a reciprocal relationship exists between the general price level and the purchasing power of the dollar. When the consumer price index or "cost-of-living" index goes up, the value of the dollar goes down, and vice versa. Higher prices lower the value of the dollar because more dollars are needed to buy a particular amount of goods, services, or resources. For example, if the price level doubles, the value of the dollar declines by one-half, or 50 percent.

Conversely, lower prices increase the purchasing power of the dollar because fewer dollars are needed to obtain a specific quantity of goods and services. If the price level falls by, say, one-half, or 50 percent, the purchasing power of the dollar doubles.

In equation form, the relationship looks like this:

$$\$V = 1/P$$

To find the value of the dollar $\$V$, divide 1 by the price level P expressed as an index number (in hundredths). If the price level is 1, then the value of the dollar is 1. If the price level rises to, say, 1.20, $\$V$ falls to 0.833; a 20 percent increase in the price level reduces the value of the dollar by 16.67 percent. Check your understanding of this reciprocal relationship by determining the value of $\$V$ and its percentage rise when P falls by 20 percent from $1 to 0.80.

Inflation and Acceptability In Chapter 27 we noted situations in which a nation's currency became worthless and unacceptable in exchange. These instances of runaway inflation, or *hyperinflation*, happened when the government issued so many pieces of paper currency that the purchasing power of each of those units of money was almost totally undermined. The infamous post–World War I hyperinflation in Germany is an example. In December 1919 there were about 50 billion marks in circulation. Four years later there were 496,585,345,900 billion marks in circulation! The result? The German mark in 1923 was worth an infinitesimal fraction of its 1919 value.[3]

Runaway inflation may significantly depreciate the value of money between the time it is received and the time it is spent. Rapid declines in the value of a currency may cause it to cease being used as a medium of exchange. Businesses and households may refuse to accept paper money in exchange because they do not want to bear the loss in its value that will occur while it is in their possession. (All this despite the fact that the government says that paper currency is legal tender!) Without an acceptable domestic medium of exchange, the economy may simply revert to barter. Alternatively, more stable currencies such as the U.S. dollar or European euro may come into widespread use. At the extreme, a country may adopt a foreign currency as its own official currency as a way to counter hyperinflation.

Similarly, people will use money as a store of value only as long as there is no sizable deterioration in the value of that money because of inflation. And an economy can effectively employ money as a unit of account only when its purchasing power is relatively stable. A monetary yardstick that no longer measures a yard (in terms of purchasing power) does not permit buyers and sellers to establish the terms of trade clearly. When the value of the dollar is declining rapidly, sellers do not know what to charge and buyers do not know what to pay.

[3]Frank G. Graham, *Exchange, Prices, and Production in Hyperinflation Germany, 1920–1923* (Princeton, N.J.: Princeton University Press, 1930), p. 13.

Stabilizing Money's Purchasing Power

Rapidly rising price levels (rapid inflation) and the consequent erosion of the purchasing power of money typically result from imprudent economic policies. Since the purchasing power of money and the price level vary inversely, stabilization of the purchasing power of a nation's money requires stabilization of the nation's price level. Such price-level stability (2 to 3 percent annual inflation) mainly necessitates intelligent management or regulation of the nation's money supply and interest rates *(monetary policy)*. It also requires appropriate *fiscal policy* supportive of the efforts of the nation's monetary authorities to hold down inflation. In the United States, a combination of legislation, government policy, and social practice inhibits imprudent expansion of the money supply that might jeopardize money's purchasing power. The critical role of the U.S. monetary authorities (the Federal Reserve) in maintaining the purchasing power of the dollar is the subject of Chapter 34. For now, simply note that they make available a particular quantity of money, such as *M*2 in Figure 32.1, and can change that amount through their policy tools.

QUICK REVIEW 32.2

- In the United States, all money consists essentially of the debts of government, commercial banks, and thrift institutions.
- These debts efficiently perform the functions of money as long as their value, or purchasing power, is relatively stable.
- The value of money is rooted not in specified quantities of precious metals but in the amounts of goods, services, and resources that money will purchase.
- The value of the dollar (its domestic purchasing power) is inversely related to the price level.
- Government's responsibility in stabilizing the purchasing power of the monetary unit calls for (a) effective control over the supply of money by the monetary authorities and (b) the application of appropriate fiscal policies by the president and Congress.

The Federal Reserve and the Banking System

LO32.4 Discuss the makeup of the Federal Reserve and its relationship to banks and thrifts.

In the United States, the "monetary authorities" we have been referring to are the members of the Board of Governors of the **Federal Reserve System** (the "Fed"). As shown in Figure 32.2, the Board directs the activities of the 12 Federal Reserve Banks, which in turn control the lending activity of the nation's banks and thrift institutions. The Fed's major goal is to control the money supply. But since checkable deposits in banks are such a large part of the money supply, an important part of its duties involves assuring the stability of the banking system.

Historical Background

Early in the twentieth century, Congress decided that centralization and public control were essential for an efficient banking system. Decentralized, unregulated banking had fostered the inconvenience and confusion of numerous private bank notes being used as currency. It also had resulted in occasional episodes of monetary mismanagement such that the money supply was inappropriate to the needs of the economy. Sometimes "too much" money precipitated rapid inflation; other times "too little money" stunted the economy's growth by hindering the production and exchange of goods and services. No single entity was charged with creating and implementing nationally consistent banking policies.

Furthermore, acute problems in the banking system occasionally erupted when banks either closed down or insisted on immediate repayment of loans to prevent their own failure. At such times, a banking crisis could emerge, with individuals and businesses who had lost confidence in their banks attempting to simultaneously withdraw all of their money—thereby further crippling the already weakened banks.

An unusually acute banking crisis in 1907 motivated Congress to appoint the National Monetary Commission to study the monetary and banking problems of the economy and to outline a course of action for Congress. The result was the Federal Reserve Act of 1913.

Let's examine the various parts of the Federal Reserve System and their relationship to one another.

Board of Governors

The central authority of the U.S. money and banking system is the **Board of Governors** of the Federal Reserve System. The U.S. president, with the confirmation of the Senate, appoints the seven Board members. Terms are 14 years and staggered so that one member is replaced every 2 years. In addition, new members are appointed when resignations occur. The president selects the chairperson and vice chairperson of the Board from among the members. Those officers serve 4-year terms and can be reappointed to new 4-year terms by the president. The long-term appointments provide the Board with continuity, experienced membership, and independence from political pressures that could result in inflation.

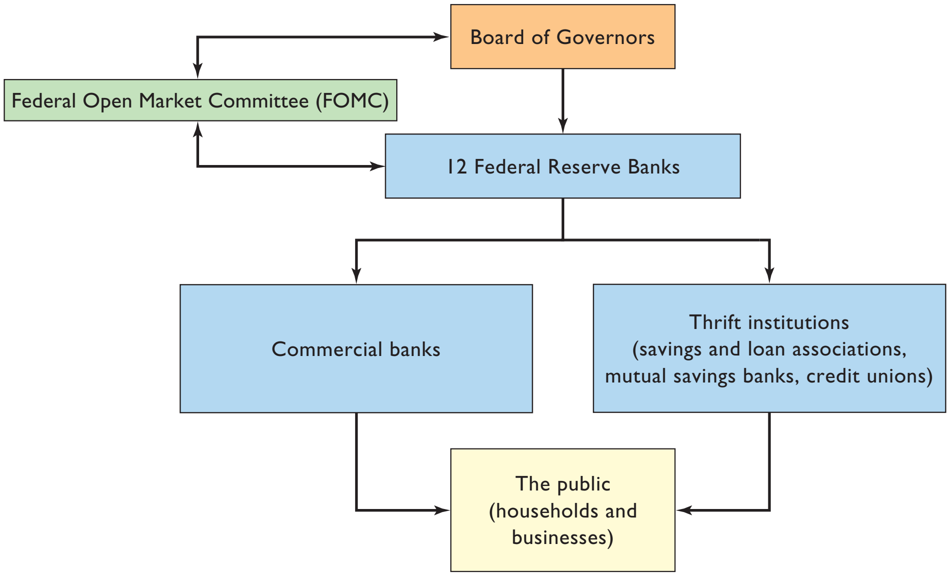


FIGURE 32.2 Framework of the Federal Reserve System and its relationship to the public. The Board of Governors makes the basic policy decisions that provide monetary control of the U.S. money and banking systems. The 12 Federal Reserve Banks implement these decisions. Both the Board of Governors and the 12 Federal Reserve Banks are aided by the Federal Open Market Committee (FOMC).

The 12 Federal Reserve Banks

The 12 **Federal Reserve Banks,** which blend private and public control, collectively serve as the nation's "central bank." These banks also serve as bankers' banks.

Central Bank Most nations have a single central bank—for example, Britain's Bank of England or Japan's Bank of Japan. The United States' central bank consists of 12 banks whose policies are coordinated by the Fed's Board of Governors. The 12 Federal Reserve Banks accommodate the geographic size and economic diversity of the United States and the nation's large number of commercial banks and thrifts.

Figure 32.3 locates the 12 Federal Reserve Banks and indicates the district that each serves. These banks implement the basic policy of the Board of Governors.

Quasi-Public Banks The 12 Federal Reserve Banks are quasi-public banks, which blend private ownership and public control. Each Federal Reserve Bank is owned by the private commercial banks in its district. (Federally chartered banks are required to purchase shares of stock in

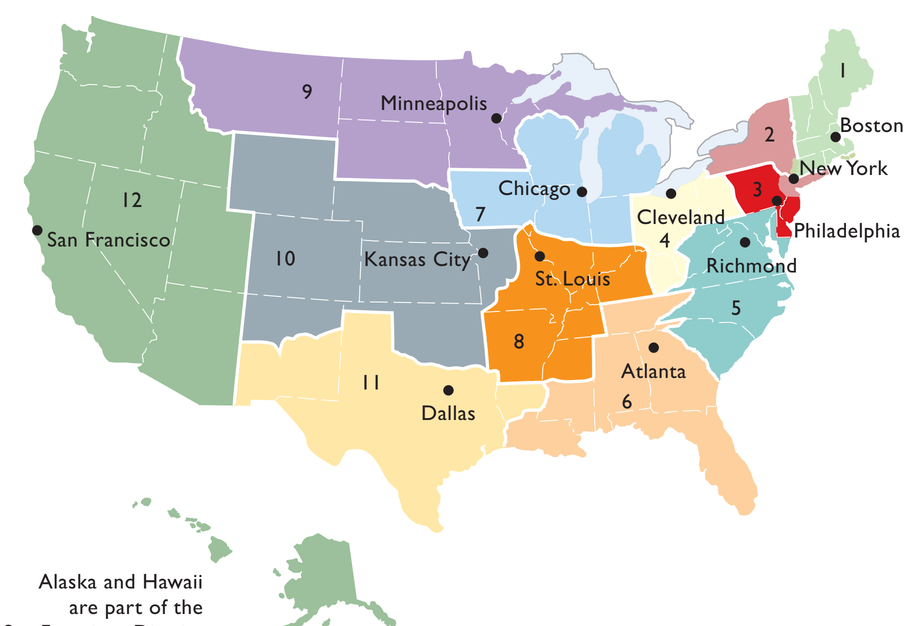


FIGURE 32.3 The 12 Federal Reserve Districts. The Federal Reserve System divides the United States into 12 districts, each having one central bank and in some instances one or more branches of the central bank.

Source: Federal Reserve Bulletin, **www.federalreserve.gov/pubs/bulletin**.

the Federal Reserve Bank in their district.) But the Board of Governors, a government body, sets the basic policies that the Federal Reserve Banks pursue.

Despite their private ownership, the Federal Reserve Banks are in practice public institutions. Unlike private firms, they are not motivated by profit. The policies they follow are designed by the Board of Governors to promote the well-being of the economy as a whole. Thus, the activities of the Federal Reserve Banks are frequently at odds with the profit motive.[4] Also, the Federal Reserve Banks do not compete with commercial banks. In general, they do not deal with the public; rather, they interact with the government and commercial banks and thrifts.

Bankers' Banks The Federal Reserve Banks are "bankers' banks." They perform essentially the same functions for banks and thrifts as those institutions perform for the public. Just as banks and thrifts accept the deposits of and make loans to the public, so the central banks accept the deposits of and make loans to banks and thrifts. Normally, these loans average only about $150 million a day. But in emergency circumstances the Federal Reserve Banks become the "lender of last resort" to the banking system and can lend out as much as needed to ensure that banks and thrifts can meet their cash obligations. On the day after terrorists attacked the United States on September 11, 2001, the Fed lent $45 *billion* to U.S. banks and thrifts. The Fed wanted to make sure that the destruction and disruption in New York City and the Washington, D.C., area did not precipitate a nationwide banking crisis.

The Fed assumed an even greater role as a lender of last resort during the financial crisis of 2007–2008. We discuss that crisis and the Fed's emergency response to the crisis later in this chapter.

But the Federal Reserve Banks have a third function, which banks and thrifts do not perform: They issue currency. Congress has authorized the Federal Reserve Banks to put into circulation Federal Reserve Notes, which constitute the economy's paper money supply.

FOMC

The **Federal Open Market Committee (FOMC)** aids the Board of Governors in conducting monetary policy. The FOMC is made up of 12 individuals:

- The seven members of the Board of Governors.
- The president of the New York Federal Reserve Bank.
- Four of the remaining presidents of Federal Reserve Banks on a 1-year rotating basis.

The FOMC meets regularly to direct the purchase and sale of government securities (bills, notes, bonds) in the open market in which such securities are bought and sold on a daily basis. We will find in Chapter 34 that the purpose of these aptly named *open-market operations* is to control the nation's money supply and influence interest rates. The Federal Reserve Bank in New York City conducts most of the Fed's open-market operations.

Commercial Banks and Thrifts

There are about 6,000 commercial banks. Roughly three-fourths are state banks. These are private banks chartered (authorized) by the individual states to operate within those states. One-fourth are private banks chartered by the federal government to operate nationally; these are national banks. Some of the U.S. national banks are very large, ranking among the world's largest financial institutions (see Global Perspective 32.1).

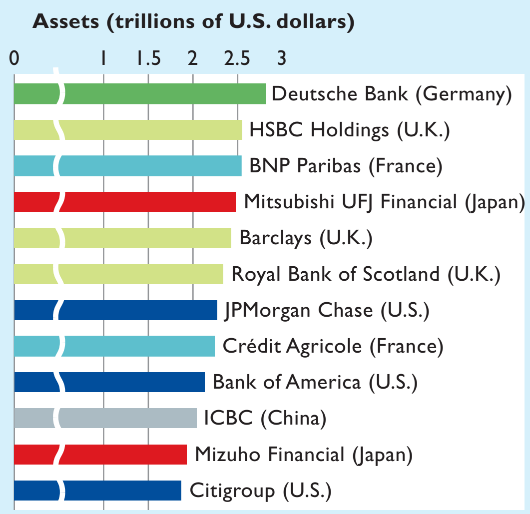
The 8,500 thrift institutions—most of which are credit unions—are regulated by agencies in addition to the

[4]Although it is not their goal, the Federal Reserve Banks have actually operated profitably, largely as a result of the Treasury debts they hold. Part of the profit is used to pay 6 percent annual dividends to the commercial banks that hold stock in the Federal Reserve Banks; the remaining profit is usually turned over to the U.S. Treasury.

GLOBAL PERSPECTIVE 32.1

The World's 12 Largest Financial Institutions

The world's 12 largest private sector financial institutions are headquartered in Europe, Japan, and the United States (2012 data).



Source: **www.forbes.com/global2000/list/**. Reprinted by Permission of Forbes Media LLC © 2012.

Board of Governors and the Federal Reserve Banks. For example, credit unions are regulated and monitored by the National Credit Union Administration (NCUA). But the thrifts *are* subject to monetary control by the Federal Reserve System. In particular, like the banks, thrifts are required to keep a certain percentage of their checkable deposits as "reserves." In Figure 32.2 we use arrows to indicate that the thrift institutions are subject to the control of the Board of Governors and the central banks. Decisions concerning monetary policy affect the thrifts along with the commercial banks.

Fed Functions, Responsibilities, and Independence

LO32.5 Identify the functions and responsibilities of the Federal Reserve and explain why Fed independence is important.

The Fed performs several functions, some of which we have already identified, but they are worth repeating:

- ***Issuing currency*** The Federal Reserve Banks issue Federal Reserve Notes, the paper currency used in the U.S. monetary system. (The Federal Reserve Bank that issued a particular bill is identified in black in the upper left of the front of the newly designed bills. "A1," for example, identifies the Boston bank, "B2" the New York bank, and so on.)
- ***Setting reserve requirements and holding reserves*** The Fed sets reserve requirements, which are the fractions of checking account balances that banks must maintain as currency reserves. The central banks accept as deposits from the banks and thrifts any portion of their mandated reserves not held as vault cash.
- ***Lending to financial institutions and serving as an emergency lender of last resort*** The Fed makes routine short-term loans to banks and thrifts and charges them an interest rate called the *discount rate*. It also occasionally auctions off loans to banks and thrifts through its *Term Auction Facility*, discussed in Chapter 34. In times of financial emergencies, the Fed serves as a lender of last resort to critical parts of the U.S. financial industry.
- ***Providing for check collection*** The Fed provides the banking system with a means for collecting on checks. If Sue writes a check on her Miami bank or thrift to Joe, who deposits it in his Dallas bank or thrift, how does the Dallas bank collect the money represented by the check drawn against the Miami bank? Answer: The Fed handles it by adjusting the reserves (deposits) of the two banks.
- ***Acting as fiscal agent*** The Fed acts as the fiscal agent (provider of financial services) for the federal government. The government collects huge sums through taxation, spends equally large amounts, and sells and redeems bonds. To carry out these activities, the government uses the Fed's facilities.
- ***Supervising banks*** The Fed supervises the operations of banks. It makes periodic examinations to assess bank profitability, to ascertain that banks perform in accordance with the many regulations to which they are subject, and to uncover questionable practices or fraud. Following the financial crisis of 2007–2008, Congress expanded the Fed's supervisory powers over banks.[5]
- ***Controlling the money supply*** Finally, the Fed has ultimate responsibility for regulating the supply of money, and this enables it to influence interest rates. The major task of the Fed under usual economic circumstances is to manage the money supply (and thus interest rates) according to the needs of the economy. This involves making an amount of money available that is consistent with high and rising levels of output and employment *and* a relatively stable price level. While most of the other functions of the Fed are routine activities or have a service nature, managing the nation's money supply requires making basic, but unique, policy decisions. (We discuss those decisions in detail in Chapter 34.)

Federal Reserve Independence

Congress purposely established the Fed as an independent agency of government. The objective was to protect the Fed from political pressures so that it could effectively control the money supply and maintain price stability. Political pressures on Congress and the executive branch may at times result in inflationary fiscal policies, including tax cuts and special-interest spending. If Congress and the executive branch also controlled the nation's monetary policy, citizens and lobbying groups undoubtedly would pressure elected officials to keep interest rates low even

[5]The Fed is not alone in this task of supervision. The individual states supervise the banks that they charter. The Office of the Comptroller of the Currency has separate supervisory authority over the banks and the thrifts. Also, the Federal Deposit Insurance Corporation supervises the banks and thrifts whose deposits it insures.

though at times high interest rates are necessary to reduce aggregate demand and thus control inflation. An independent monetary authority (the Fed) can take actions to increase interest rates when higher rates are needed to stem inflation. Studies show that countries that have independent central banks like the Fed have lower rates of inflation, on average, than countries that have little or no central bank independence.

QUICK REVIEW 32.3

- The U.S. banking system consists of (a) the Board of Governors of the Federal Reserve System, (b) the 12 Federal Reserve Banks, and (c) some 6,000 commercial banks and 8,500 thrift institutions (mainly credit unions).
- The 12 Federal Reserve Banks are simultaneously (a) central banks, (b) quasi-public banks, and (c) bankers' banks.
- The major functions of the Fed are to (a) issue Federal Reserve Notes, (b) set reserve requirements and hold reserves deposited by banks and thrifts, (c) lend money to financial institutions and serve as the lender of last resort in national financial emergencies, (d) provide for the rapid collection of checks, (e) act as the fiscal agent for the federal government, (f) supervise the operations of the banks, and (g) regulate the supply of money in the best interests of the economy.

The Financial Crisis of 2007 and 2008

LO32.6 Identify and explain the main factors that contributed to the financial crisis of 2007–2008.

As previously noted, a properly functioning monetary system supports the continuous circular flows of income and expenditures in the economy. In contrast, a malfunctioning monetary system causes major problems in credit markets and can cause severe fluctuations in the economy's levels of output, employment, and prices.

"Malfunctioning" is too gentle an adjective to describe the monetary system in late 2007 and 2008. In that period, the U.S. financial system faced its most serious crisis since the Great Depression of the 1930s. The financial crisis soon spread to the entire economy, culminating in the severe recession of 2007–2009. We discussed the recession in detail in previous chapters, and we now want to examine the financial crisis that led up to it. What was the nature of the financial crisis? What caused it? How has it changed the structure of the U.S. financial services industry?

The Mortgage Default Crisis

In 2007 a major wave of defaults on home mortgage loans threatened the health of not only the original mortgage lenders but of any financial institution that had made such loans or invested in such loans either directly or indirectly. A majority of these mortgage defaults were on **subprime mortgage loans**—high-interest-rate loans to home buyers with higher-than-average credit risk. Ironically, the federal government had encouraged banks to make these types of loans as part of an effort to broaden home ownership to more Americans. But more directly to the point, several of the biggest indirect investors in these subprime loans had been banks. The banks had lent money to investment companies that had purchased many of the mortgages from mortgage lenders. When the mortgages started to go bad, many investment funds "blew up" and could not repay the loans they had taken out from the banks. The banks thus had to "write off" (declare unrecoverable) the loans they had made to the investment companies, but doing that meant reducing their banks' reserves and limiting their ability to generate new loans. This greatly threatened the economy because both consumers and businesses rely on loans to finance consumption and investment expenditures.

A strange thing about the crisis was that before it happened, banks and government regulators had mistakenly believed that an innovation known as the "mortgage-backed security" had eliminated most of the bank exposure to mortgage defaults. **Mortgage-backed securities** are bonds backed by mortgage payments. To create them, banks and other mortgage lenders first made mortgage loans. But then instead of holding all of those loans as assets on their balance sheets and collecting the monthly mortgage payments, the banks and other mortgage lenders bundled hundreds or thousands of them together and sold them off as bonds—in essence selling the right to collect all the future mortgage payments. The banks obtained a single, up-front cash payment for the bond and the bond buyer started to collect the mortgage payments as the return on the investment.

From the banks' perspective, this seemed like a smart business decision because it transferred any future default risk on those mortgages to the buyer of the bond. The banks thought that they were off the hook for these mortgages. Unfortunately for them, however, they lent a substantial portion of the money they received from selling the bonds to investment funds that invested in mortgage-backed bonds. They also purchased large amounts of mortgage-backed securities as financial investments to help meet bank capital requirements set by bank regulators. So while the banks were no longer directly exposed to major portions of the mortgage default risk, they were still indirectly exposed to it. When many homebuyers started to default on their mortgages, the banks lost money on the mortgages they still held. The banks also lost money on the loans they had made to the investors who had purchased mortgage-backed securities, and also on the mortgage-backed securities the banks had purchased from investment firms.

But what had caused the skyrocketing mortgage default rates in the first place? There were many causes, including certain government programs that greatly encouraged and subsidized home ownership for former renters. Also contributing were declining real estate values that arrived at the end of a long housing boom during which house prices had greatly increased. But an equally important factor was the bad incentives provided by the previously discussed mortgage-backed bonds. Because the banks and other mortgage lenders thought that they were no longer exposed to large portions of their mortgage default risk, they became lax in their lending practices—so much so that people were granted subprime mortgage loans that they were unlikely to be able to repay. Some mortgage companies were so eager to sign up new homebuyers (in order to bundle their loans together to sell bonds) that they stopped running credit checks and even allowed applicants to claim higher incomes than they were actually earning in order to qualify them for big loans. The natural result was that many people took on "too much mortgage" and were soon failing to make their monthly payments.

Securitization

The problems just described relate to **securitization**—the process of slicing up and bundling groups of loans, mortgages, corporate bonds, or other financial debts into distinct new securities. This process was not new and was viewed favorably by government regulators, who thought securitization made the banking system safer by allowing banks to shed risk. As noted in our discussion of mortgages, these securities were sold to financial investors, who purchased them to obtain the interest payments and the eventual return of principal generated by the underlying securities. For example, the mortgage loans provided to the subprime borrowers were bundled together as mortgage-backed securities and sold to private investors, mutual fund firms, and pension funds. These securities were attractive to many private investors and financial institutions alike because they offered higher-interest returns than securities backed by less-risky mortgages or other safer investments.

Once created, loan-backed securities are bought and sold in financial markets just like other securities such as stocks and bonds. These sorts of securities can therefore end up worldwide in the investment portfolios of banks, thrifts, insurance companies, and pensions, as well as in personal accounts.

To reduce the risk for holders of these securities, a few large insurance companies developed other securities that the holders of loan-backed securities could purchase to insure against losses from defaults. American International Group (AIG), in particular, issued billions of dollars of *collateralized default swaps*—essentially insurance policies—that were designed to compensate the holders of loan-backed securities if the loans underlying these investments went into default and did not pay off. Thus, collateralized default swaps became yet another category of investment security that was highly exposed to mortgage-loan risk.

Securitization is so widespread and so critical to the modern financial system that economists sometimes refer to it as the *shadow banking system*. All sorts of securities backed by loans or other securities are issued, bought, sold, and resold each day in a process that helps to keep credit flowing to the households and firms that rely on it for their personal and business needs. In general, securitization therefore is a positive financial innovation. But mortgage-backed securities, in particular, turned out to contain much more risk than most people thought.

Investors and government regulators failed to ask three related questions about mortgage-backed securities: What would happen if the value of one of the types of loans (say, mortgages) that underlies part of the securitization process unexpectedly plunged? And what then would happen if some of the largest holders of the securities based on these mortgages were major U.S. financial institutions that are vitally important to the day-to-day financing of the credit needed to keep the American economy running smoothly? And what would happen after that if the main insurer of these securities not only was the largest insurance company in the United States but in the world?

All three seemingly improbable "what ifs?" occurred! As previously explained, interest rates on adjustable-rate mortgages increased and house prices fell. Borrowers who had made relatively small down payments on home purchases or had previously cashed out home equity through refinancing discovered that they owed more on their mortgages than their properties were worth. Their loans were said to be "underwater." As interest rates adjusted upward and the economy slowed, borrowers began falling behind on their monthly mortgage payments. Lenders began to foreclose on many houses, while other borrowers literally handed in their house keys and walked away from their houses *and* their mortgages.

Failures and Near-Failures of Financial Firms

When the mortgage loan "card" underpinning mortgage-based securitization fell, the securitization layers above it collapsed like a house of cards. First, the big mortgage lenders faced demise because they still held large amounts of the bad debt. Three huge mortgage lenders collapsed or nearly collapsed. Countrywide, the second largest mortgage lender, was saved from bankruptcy by Bank of America. Regulators also seized Washington Mutual bank, the nation's largest mortgage lender, and arranged a quick takeover by JPMorgan Chase. Wachovia bank's heavy exposure to mortgages through its Golden West subsidiary resulted in near bankruptcy, and it was rescued through acquisition by Wells Fargo.

The exposure to the growing problem of loan defaults quickly jumped from direct mortgage lenders to other financial institutions. Securities firms and investment banks that held large amounts of loan-backed securities began to suffer huge losses. Merrill Lynch lost more in two years than it made in the prior decade and was acquired at a fire-sale price by Bank of America. Lehman Brothers, a major holder of mortgage-backed securities, declared bankruptcy. Goldman Sachs, Morgan Stanley, and other financial firms that had heavy exposures to mortgage-backed securities and collateralized default swaps rushed to become bank holding companies so they could qualify for the massive emergency loans that the Federal Reserve was making available to banks and bank holding companies. Citibank survived through infusions of federal government loans. Insurance company AIG suffered enormous losses because it had not set aside sufficient reserves to pay off the unexpectedly large losses that accrued on the insurance policies that it had sold to holders of mortgage-backed securities. The nightmarish thought of a total collapse of the U.S. financial system suddenly became a realistic possibility.

QUICK REVIEW 32.4

- The financial crisis of 2007–2008 consisted of an unprecedented rise in mortgage loan defaults, the collapse or near collapse of several major financial institutions, and the generalized freezing up of credit availability.
- The crisis resulted from bad mortgage loans together with declining real estate prices.
- The crisis exposed the underestimation of risk by holders of mortgage-backed securities as well as faulty insurance securities that had been designed to protect holders of mortgage-backed securities from the risk of default.

The Policy Response to the Financial Crisis

LO32.7 Discuss the actions of the U.S. Treasury and the Federal Reserve that helped keep the banking and financial crisis of 2007–2008 from worsening.

The U.S. government responded to the financial crisis with historically unprecedented fiscal policy actions while the Fed acted aggressively as a lender of last resort.

The Treasury Bailout: TARP

In late 2008 Congress passed the **Troubled Asset Relief Program (TARP),** which allocated $700 billion—yes, billion—to the U.S. Treasury to make emergency loans to critical financial and other U.S. firms. Most of this "bailout" money eventually was lent out. In fact, as of March 2009, the federal government and Federal Reserve had spent $170 billion just keeping insurer AIG afloat. Other major recipients of TARP funds included Citibank, Bank of America, JPMorgan Chase, and Goldman Sachs. Later, nonfinancial firms such as General Motors and Chrysler also received several billion dollars of TARP loans.

TARP indeed saved several financial institutions whose bankruptcy would have caused a tsunami of secondary effects that probably would have brought down other financial firms and frozen credit throughout the economy. But this very fact demonstrates the problem of **moral hazard.** As it relates to financial investment, moral hazard is the tendency for financial investors and financial services firms to take on greater risks because they assume they are at least partially insured against losses. Without TARP, several firms would have gone bankrupt and their stockholders, bondholders, and executives all would have suffered large personal losses. With TARP, those outcomes were at least partially avoided. TARP and similar government bailouts were essentially government-provided insurance payouts to financial firms that never had to pay a single cent in insurance premiums for the massive bailouts that kept them afloat.

The correct assumption by large firms that they were simply too big for government to let them fail may have given them an incentive to make riskier investments than if no government bailouts were likely to be forthcoming.

The Fed's Lender-of-Last-Resort Activities

As noted in our previous list of Fed functions, one of the roles of the Federal Reserve is to serve as the lender of last resort to financial institutions in times of financial emergencies. The Fed performed this vital function well following the 9/11 terrorist attacks. The financial crisis of

2007–2008 presented another, broader-based financial emergency. Under Fed Chair Ben Bernanke, the Fed designed and implemented several highly creative new lender-of-last-resort facilities to pump liquidity into the financial system. These facilities, procedures, and capabilities were in addition to both the TARP efforts by the U.S. Treasury and the Fed's use of standard tools of monetary policy (the subject of Chapter 34) designed to reduce interest rates. All the new Fed facilities had the single purpose and desired outcome of keeping credit flowing.

Total Fed assets rose from $885 billion in February 2008 to $1,903 billion in March 2009. This increase reflected a huge rise in the amount of securities (U.S. securities, mortgage-backed securities, and others) owned by the Fed. In undertaking its lender-of-last-resort functions, the Fed bought these securities from financial institutions. The purpose was to increase liquidity in the financial system by exchanging illiquid bonds (that the firms could not easily sell during the crisis) for cash, the most liquid of all assets.

Many economists believe that TARP and the Fed's actions helped avert a second Great Depression. The following list of new Fed credit facilities underscores the extraordinary extent of the Fed's lender-of-last-resort response to the crisis.

- ***Primary Dealer Credit Facility (PDCF)*** Provided overnight loans to primary dealers who were willing to post loan-backed securities as collateral. (The Fed kept the collateral on any loan that was not repaid on time.) Primary dealers are the 21 major financial institutions that the Fed uses to buy and sell U.S. securities.
- ***Term Securities Lending Facility (TSLF)*** Lent U.S. securities to primary dealers for one-month terms to promote liquidity in the markets for these U.S. securities. The financial institutions obtained the securities from the Fed through participating in competitive single-bid auctions.
- ***Asset-Backed Commercial Paper Money Market Mutual Fund Liquidity Facility*** Provided loans to U.S. banks and thrifts to finance their purchases of *commercial paper* from money market mutual funds. Commercial paper consists of asset-backed, short-term IOUs that are mainly issued by corporations. These short-term loans are vital for financing the day-to-day operations of businesses.
- ***Commercial Paper Funding Facility (CPFF)*** Purchased commercial paper to support the commercial paper market and therefore the short-term credit needs of businesses.
- ***Money Market Investor Funding Facility (MMIFF)*** Provided funding support to a private-sector initiative designed to ensure the liquidity of U.S. money market mutual funds. Many Americans rely on money market mutual funds as low-risk investments.
- ***Term Asset-Backed Securities Loan Facility (TALF)*** Helped households and business with their credit needs by providing funding support for asset-backed securities collateralized by student loans, auto loans, credit card loans, and loans guaranteed by the Small Business Administration (SBA).
- ***Interest Payments on Reserves*** Bolstered the profitability of banks by paying interest on the reserves they hold in their vaults or in the Federal Reserve Banks.

These extraordinary efforts, like those of the Treasury, helped prevent total disarray in the credit markets. But like TARP, the Fed efforts intensified the moral hazard problem by greatly limiting the losses that otherwise would have resulted from bad financial assumptions and decisions.

QUICK REVIEW 32.5

- The Troubled Asset Relief Program (TARP) authorized the U.S. Treasury to spend up to $700 billion to make emergency loans and guarantees to failing financial firms.
- The Treasury rescue, or bailout, was aided by lender-of-last-resort loans provided by the Federal Reserve to financial institutions through a series of newly established Fed facilities.
- The TARP loans and the Fed's lender-of-last-resort actions intensified the moral hazard problem in which financial investors and financial firms take on greater risk because they assume that the government will bail them out if they lose money.

The Postcrisis U.S. Financial Services Industry

LO32.8 Identify the main subsets of the financial services industry in the United States and provide examples of some firms in each category.

Table 32.1 lists the major categories of firms within the U.S. financial services industry and gives examples of firms in each category. Note that the main categories of the **financial services industry** are commercial banks, thrifts,

LAST WORD

Too Big to Fail. Too Big to Jail?

In 2010, Congress Passed the Wall Street Reform and Consumer Protection Act in an Attempt to Better Regulate Big Financial Firms. But Their Enormous Size Has Now, Apparently, Placed Them Above the Law.

In early December 2012, Lanny Breuer, the Assistant U.S. Attorney General in charge of prosecuting financial crimes, explained at a press conference that he had decided to only fine HSBC bank $1.9 billion—or about five weeks' profits—as a punishment for nearly a decade of laundering money for the Sinaloa drug cartel, Al Qaeda, and Russian mobsters; helping Iran, the Sudan, and North Korea evade sanctions; and helping hundreds of individuals and businesses cheat on their taxes. Not a single HSBC official would face jail time or a personal fine.

Breuer explained that he chose not to criminally prosecute the bank and put any of its officers in jail because "HSBC would almost certainly have lost its banking license in the United States. The future of the institution would have been under threat and the entire banking system would have been destabilized." In other words, HSBC was not prosecuted because it was very large and very interconnected with other financial firms.

A week later Breuer was again in front of the press. This time it was to announce a set of minor penalties in the LIBOR interest-rate scandal, which involved major international banks illegally manipulating the London Interbank Offer Rate (LIBOR) for their own personal benefit. That manipulation affected trillions of dollars of financial contracts around the world, including millions of home mortgages in the United States. It was thus the most extensive financial crime in world history. But

Breuer worried aloud about global financial stability and stated, "Our goal here is: Not to destroy a major financial institution."

Breuer's decision to put financial stability ahead of prosecuting financial crimes was made even more clear in early January 2013 when on the TV show *Frontline* he was asked whether prosecutors should think about anything other than pursuing justice. He responded, "Well, I think I am pursuing justice. And I think the entire responsibility of the Department is to pursue justice. But in any given case, I think I and prosecutors around

insurance companies, mutual fund companies, pension funds, security firms, and investment banks. Even before the financial crisis of 2007–2008, the financial services industry was consolidating into fewer, larger firms, each offering a wider spectrum of services. In 1999 Congress ended the Depression-era prohibition against banks selling stocks, bonds, and mutual funds. Thus, the lines between the subsets of the financial industry began to blur. Many banks acquired stock brokerage firms and, in a few cases, insurance companies. For example, Citigroup, which was once only into banking, now owns Smith Barney, a large securities firm. Many large banks (for example, Wells Fargo) and pension funds (for example, TIAA-CREF) now provide mutual funds, including money market mutual funds that pay relatively high interest and on which checks of $500 or more can be written.

The upheaval in the financial markets caused by the financial crisis of 2007–2008 further consolidated the industry and further blurred the lines between its segments. Between September 2007 and September 2009, the FDIC shut down more than 200 U.S. banks and transferred their bank deposits to other, usually larger, banks. In 2009 the three largest U.S. banks (JPMorgan Chase, Bank of America, and Wells Fargo) held roughly $3 of every $10 on deposit in the United States.

Also, during the financial crisis of 2007–2008, major investment banks Goldman Sachs and Morgan Stanley opted to become commercial banks to gain access to emergency Federal Reserve loans. The nation's largest thrift—Washington Mutual—was absorbed by commercial bank JPMorgan Chase. But even with all this blending, the categories in Table 32.1 remain helpful. The

the country, being responsible, should speak to regulators, should speak to experts, because if I bring a case against institution A, and as a result of bringing that case, there's some huge economic effect—if it creates a ripple effect so that, suddenly, counterparties and other financial institutions or other companies that had nothing to do with this are affected badly—it's a factor we need to know and understand." It is perhaps not surprising that *Frontline* titled that episode "The Untouchables."

Many economists, however, sympathize with Breuer's position because legal changes made during the 1990s resulted in a financial system dominated by just a few large firms, each of which is now so large and so interconnected that if any were to collapse, the whole system would indeed be destabilized.

Before those legal changes were made, the structure of the U.S. financial system had been defined by the Glass-Steagall Act of 1933. That law required Wall Street to segregate high-risk and low-risk financial activities across different firms. In particular, Glass-Steagall had required commercial banks to only engage in low-risk lending activities like making home mortgages and small business loans. High-risk financial activities like stock picking, derivatives trading, and investment banking had to be handled by an entirely different set of firms. Thus, even if the firms that took high risks went bankrupt, they were not going to be able to affect the traditional banking system that the entire economy depended upon for making payments, providing cash, and issuing loans to individuals and small businesses.

The legal changes made during the 1990s were aimed at allowing financial firms to offer "one-stop shopping" for financial services. Instead of customers having to go to one firm for a checking account, another to buy life insurance, and yet another to invest in mutual funds, single companies were allowed to offer every imaginable financial service. Not only was this intended to be more convenient, it was argued that it would actually increase financial stability because large firms that offered a full range of financial services would be diversified across different business activities. For example, if a firm's insurance division was losing money, some other division would probably be doing well enough to either partly or fully offset the loss.

By the time the financial crisis hit in 2007, however, things had turned out quite differently. The convenience factor was there, but the ability to go into other lines of business had caused a massive consolidation of banks, insurance companies, stock brokerages, and derivatives-trading companies. In particular, the major commercial banks were now engaged in massive amounts of speculative investing—much of it financed with the money deposited into checking accounts. Thus, when the crisis hit, it was impossible for any of the major financial firms to go bankrupt without severely affecting both the payments system and the financial viability of other financial firms. Each firm was now so large and so interconnected that it had to be treated as "too big to fail" by government regulators.

What Breuer's comments in 2012 and 2013 indicated, though, was that at least some of the firms had also become "too big to jail." There are, consequently, many economists calling for a return to a financial system in which the big financial firms are broken up into entities that are each small enough to fail without catastrophically affecting the entire financial system and in which Glass-Steagall-style separations once again divide high-risk and low-risk financial activities into separate firms.

An attempt to put that separation into practice was passed in 2010 as part of the Wall Street Reform and Consumer Protection Act. It was known as the Volker Rule after former Federal Reserve Chairman Paul Volker. But as of early 2013, it had yet to be implemented.

main lines of a firm's businesses often are in one category or another. For example, even though Goldman Sachs is licensed and regulated as a bank, it is first and foremost an investment company. And the insurance companies and pension funds do most of their business as such.

The financial crisis of 2007–2008 generated much introspection about what went wrong and how to prevent anything like it from happening again. Politicians and financial regulators tightened lending rules to offset the "pass the buck" incentives created by mortgage-backed securities and prevent loans from being issued to people who are unlikely to be able to make the required monthly payments. They also passed legislation to help homeowners who were "underwater" on mortgage loans remain in their homes.

In mid-2010 Congress passed and the president signed the **Wall Street Reform and Consumer Protection Act.** This sweeping law includes provisions that:

- Eliminate the Office of Thrift Supervision and give broader authority to the Federal Reserve to regulate all large financial institutions.
- Create a Financial Stability Oversight Council to be on the lookout for risks to the financial system.
- Establish a process for the federal government to liquidate (sell off) the assets of failing non-bank financial institutions, much like the FDIC does with failing banks.
- Provide federal regulatory oversight of mortgage-backed securities and other derivatives and require that they be traded on public exchanges.

TABLE 32.1 **Major Categories of Financial Institutions within the U.S. Financial Services Industry**

Institution	Description	Examples
Commercial banks	State and national banks that provide checking and savings accounts, sell certificates of deposit, and make loans. The Federal Deposit Insurance Corporation (FDIC) insures checking and savings accounts up to $250,000.	JPMorgan Chase, Bank of America, Citibank, Wells Fargo
Thrifts	Savings and loan associations (S&Ls), mutual saving banks, and credit unions that offer checking and savings accounts and make loans. Historically, S&Ls made mortgage loans for houses while mutual savings banks and credit unions made small personal loans, such as automobile loans. Today, major thrifts offer the same range of banking services as commercial banks. The Federal Deposit Insurance Corporation and the National Credit Union Administration insure checking and savings deposits up to $250,000.	Charter One, New York Community Bank, Pentagon Federal Credit Union, Boeing Employees Credit Union (BECU)
Insurance companies	Firms that offer policies (contracts) through which individuals pay premiums to insure against some loss, say, disability or death. In some life insurance policies and annuities, the funds are invested for the client in stocks and bonds and paid back after a specified number of years. Thus, insurance sometimes has a saving or financial-investment element.	Prudential, New York Life, Northwestern Mutual, Hartford, MetLife
Mutual fund companies	Firms that pool deposits by customers to purchase stocks or bonds (or both). Customers thus indirectly own a part of a particular set of stocks or bonds, say stocks in companies expected to grow rapidly (a growth fund) or bonds issued by state governments (a municipal bond fund).	Fidelity, Vanguard, Putnam, Janus, T. Rowe Price
Pension funds	For-profit or nonprofit institutions that collect savings from workers (or from employers on their behalf) throughout their working years and then buy stocks and bonds with the proceeds and make monthly retirement payments.	TIAA-CREF, Teamsters' Union, CalPERs
Securities firms	Firms that offer security advice and buy and sell stocks and bonds for clients. More generally known as *stock brokerage firms.*	Merrill Lynch, Smith Barney, Charles Schwab
Investment banks	Firms that help corporations and governments raise money by selling stocks and bonds. They also typically offer advisory services for corporate mergers and acquisitions as well as brokerage services and advice.	Goldman Sachs, Morgan Stanley, Deutsche Bank, Nomura Securities

- Require companies selling asset-backed securities to retain a portion of those securities so the sellers share part of the risk.
- Establish a stronger consumer financial protection role for the Fed through creation of the Bureau of Consumer Financial Protection.

Proponents of the new law say that it will help prevent many of the practices that led up to the financial crisis of 2007–2008. They also contend that the law will send a strong message to stockholders, bondholders, and executives of large financial firms that they will suffer unavoidable and extremely high personal financial losses if they allow their firms to ever again get into serious financial trouble.

Skeptics of the new law say that regulators already had all the tools they needed to prevent the financial crisis. They also point out that the government's own efforts to promote home ownership, via quasi-government institutions that purchased mortgage-backed securities, greatly contributed to the financial crisis. Critics of the new law say that it will simply impose heavy new regulatory costs on the financial industry while doing little to prevent future government bailouts. This chapter's Last Word considers those suspicions.

QUICK REVIEW 32.6

- The main categories of the U.S. financial services industry are commercial banks, thrifts, insurance companies, mutual fund companies, pension funds, securities firms, and investment banks.
- The reassembly of the wreckage from the financial crisis of 2007–2008 has further consolidated the already-consolidating financial services industry and has further blurred some of the lines between the subsets of the industry.
- The Wall Street Reform and Consumer Financial Protection Act of 2010 responded to the financial crisis by consolidating financial regulation, providing federal oversight of mortgage-backed securities, and creating the Bureau of Consumer Financial Protection.

SUMMARY

LO32.1 Identify and explain the functions of money.

Anything that is accepted as (*a*) a medium of exchange, (*b*) a unit of monetary account, and (*c*) a store of value can be used as money.

LO32.2 List and describe the components of the U.S. money supply.

There are two major definitions of the money supply. *M*1 consists of currency and checkable deposits; *M*2 consists of *M*1 plus savings deposits, including money market deposit accounts, small-denominated (less than $100,000) time deposits, and money market mutual fund balances held by individuals.

LO32.3 Describe what "backs" the money supply, making us willing to accept it as payment.

Money represents the debts of government and institutions offering checkable deposits (commercial banks and thrift institutions) and has value because of the goods, services, and resources it will command in the market. Maintaining the purchasing power of money depends largely on the government's effectiveness in managing the money supply.

LO32.4 Discuss the makeup of the Federal Reserve and its relationship to banks and thrifts.

The U.S. banking system consists of (*a*) the Board of Governors of the Federal Reserve System, (*b*) the 12 Federal Reserve Banks, and (*c*) some 6,000 commercial banks and 8,500 thrift institutions (mainly credit unions). The Board of Governors is the basic policymaking body for the entire banking system. The directives of the Board and the Federal Open Market Committee (FOMC) are made effective through the 12 Federal Reserve Banks, which are simultaneously (*a*) central banks, (*b*) quasi-public banks, and (*c*) bankers' banks.

LO32.5 Identify the functions and responsibilities of the Federal Reserve and explain why Fed independence is important.

The major functions of the Fed are to (*a*) issue Federal Reserve Notes, (*b*) set reserve requirements and hold reserves deposited by banks and thrifts, (*c*) lend money to financial institutions and serve as the lender of last resort in national financial emergencies, (*d*) provide for the rapid collection of checks, (*e*) act as the fiscal agent for the federal government, (*f*) supervise the operations of the banks, and (*g*) regulate the supply of money in the best interests of the economy.

The Fed is essentially an independent institution, controlled neither by the president of the United States nor by Congress. This independence shields the Fed from political pressure and allows it to raise and lower interest rates (via changes in the money supply) as needed to promote full employment, price stability, and economic growth.

LO32.6 Identify and explain the main factors that contributed to the financial crisis of 2007–2008.

The financial crisis of 2007–2008 consisted of an unprecedented rise in mortgage loan defaults, the collapse or near-collapse of several major financial institutions, and the generalized freezing up of credit availability. The crisis resulted from bad mortgage loans together with declining real estate prices. It also resulted from underestimation of risk by holders of mortgage-backed securities and faulty insurance securities designed to protect holders of mortgage-backed securities from the risk of default.

LO32.7 Discuss the actions of the U.S. Treasury and the Federal Reserve that helped keep the banking and financial crisis of 2007–2008 from worsening.

In 2008 Congress passed the Troubled Asset Relief Program (TARP), which authorized the U.S. Treasury to spend up to $700 billion to make emergency loans and guarantees to failing financial firms. The Treasury rescue, or bailout, was aided by lender-of-last-resort loans provided by the Federal Reserve to financial institutions through a series of newly established Fed facilities.

The TARP loans and the Fed's lender-of-last-resort actions intensify the moral hazard problem. This is the tendency of financial investors and financial firms to take on greater risk when they assume they are at least partially insured against loss.

LO32.8 Identify the main subsets of the financial services industry in the United States and provide examples of some firms in each category.

The main categories of the U.S. financial services industry are commercial banks, thrifts, insurance companies, mutual fund companies, pension funds, securities firms, and investment banks. The reassembly of the wreckage from the financial crisis of 2007–2008 has further consolidated the already-consolidating financial services industry and has further blurred some of the lines between the subsets of the industry.

In response to the financial crisis, Congress passed the Wall Street Reform and Consumer Financial Protection Act of 2010.

TERMS AND CONCEPTS

medium of exchange	*M*1	commercial banks
unit of account	Federal Reserve Notes	thrift institutions
store of value	token money	near-monies
liquidity	checkable deposits	*M*2

savings account
money market deposit account (MMDA)
time deposits
money market mutual fund (MMMF)
legal tender
Federal Reserve System
Board of Governors
Federal Reserve Banks
Federal Open Market Committee (FOMC)
subprime mortgage loans
mortgage-backed securities
securitization
Troubled Asset Relief Program (TARP)
moral hazard
financial services industry
Wall Street Reform and Consumer Protection Act

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What are the three basic functions of money? Describe how rapid inflation can undermine money's ability to perform each of the three functions. **LO32.1**
2. Which two of the following financial institutions offer checkable deposits included within the *M*1 money supply: mutual fund companies; insurance companies; commercial banks; securities firms; thrift institutions? Which of the following items is not included in either *M*1 or *M*2: currency held by the public; checkable deposits; money market mutual fund balances; small-denominated (less than $100,000) time deposits; currency held by banks; savings deposits? **LO32.2**
3. What are the components of the *M*1 money supply? What is the largest component? Which of the components of *M*1 is legal tender? Why is the face value of a coin greater than its intrinsic value? What near-monies are included in the *M*2 money supply? **LO32.2**
4. Explain and evaluate the following statements: **LO32.2**
   a. The invention of money is one of the great achievements of humankind, for without it the enrichment that comes from broadening trade would have been impossible.
   b. Money is whatever society says it is.
   c. In the United States, the debts of government and commercial banks are used as money.
   d. People often say they would like to have more money, but what they usually mean is that they would like to have more goods and services.
   e. When the price of everything goes up, it is not because everything is worth more but because the currency is worth less.
   f. Any central bank can create money; the trick is to create enough, but not too much, of it.
5. What "backs" the money supply in the United States? What determines the value (domestic purchasing power) of money? How does the purchasing power of money relate to the price level? Who in the United States is responsible for maintaining money's purchasing power? **LO32.3**
6. How is the chairperson of the Federal Reserve System selected? Describe the relationship between the Board of Governors of the Federal Reserve System and the 12 Federal Reserve Banks. What is the purpose of the Federal Open Market Committee (FOMC)? What is its makeup? **LO32.4**
7. The following are two hypothetical ways in which the Federal Reserve Board might be appointed. Would you favor either of these two methods over the present method? Why or why not? **LO32.4**
   a. Upon taking office, the U.S. president appoints seven people to the Federal Reserve Board, including a chair. Each appointee must be confirmed by a majority vote of the Senate, and each serves the same 4-year term as the president.
   b. Congress selects seven members from its ranks (four from the House of Representatives and three from the Senate) to serve at congressional pleasure as the Board of Governors of the Federal Reserve System.
8. What is meant when economists say that the Federal Reserve Banks are central banks, quasi-public banks, and bankers' banks? **LO32.4**
9. Why do economists nearly uniformly support an independent Fed rather than one beholden directly to either the president or Congress? **LO32.5**
10. Identify three functions of the Federal Reserve of your choice, other than its main role of controlling the supply of money. **LO32.5**
11. How does each of the following relate to the financial crisis of 2007–2008: declines in real estate values, subprime mortgage loans, mortgage-backed securities, AIG. **LO32.6**
12. What is TARP and how was it funded? What is meant by the term "lender of last resort" and how does it relate to the financial crisis of 2007–2008? How do government and Federal Reserve emergency loans relate to the concept of moral hazard? **LO32.7**
13. What are the major categories of firms that make up the U.S. financial services industry? Are there more or fewer banks today than before the start of the financial crisis of 2007–2008? Why are the lines between the categories of financial firms even more blurred than they were before the

crisis? How did the Wall Street Reform and Consumer Protection Act of 2010 try to address some of the problems that helped cause the crisis? **LO32.8**

14. **LAST WORD** Why are federal prosecutors reluctant to bring major charges against large financial firms? What was the main regulatory action of the Glass-Steagall law? Why might having many smaller financial firms be more stable than having fewer larger firms? What argument can be made for the possibility that larger financial firms might be more stable than smaller financial firms?

REVIEW QUESTIONS

1. The three functions of money are: **LO32.1**
   - **a.** Liquidity, store of value, and gifting.
   - **b.** Medium of exchange, unit of account, and liquidity.
   - **c.** Liquidity, unit of account, and gifting.
   - **d.** Medium of exchange, unit of account, and store of value.
2. Suppose that a small country currently has $4 million of currency in circulation, $6 million of checkable deposits, $200 million of savings deposits, $40 million of small-denominated time deposits, and $30 million of money market mutual fund deposits. From these numbers we see that this small country's *M*1 money supply is ____________, while its *M*2 money supply is ____________. **LO32.2**
   - **a.** $10 million; $280 million.
   - **b.** $10 million; $270 million.
   - **c.** $210 million; $280 million.
   - **d.** $250 million; $270 million.
3. Recall the formula that states that $\$V = 1/P$, where V is the value of the dollar and P is the price level. If the price level falls from 1 to 0.75, what will happen to the value of the dollar? **LO32.3**
   - **a.** It will rise by a third (33.3 percent).
   - **b.** It will rise by a quarter (25 percent).
   - **c.** It will fall by a quarter (−25 percent).
   - **d.** It will fall by a third (−33.3 percent).
4. Which group votes on the open-market operations that are used to control the U.S. money supply and interest rates? **LO32.4**
   - **a.** The Federal Reserve System.
   - **b.** The 12 Federal Reserve Banks.
   - **c.** The Board of Governors of the Federal Reserve System.
   - **d.** The Federal Open Market Committee (FOMC).
5. An important reason why members of the Federal Reserve's Board of Governors are each given extremely long, 14-year terms is to: **LO32.4**
   - **a.** Insulate members from political pressures that could result in inflation.
   - **b.** Help older members avoid job searches before retiring.
   - **c.** Attract younger people with lots of time left in their careers.
   - **d.** Avoid the trouble of constantly having to deal with new members.
6. Which of the following is ***not*** a function of the Fed? **LO32.5**
   - **a.** Setting reserve requirements for banks.
   - **b.** Advising Congress on fiscal policy.
   - **c.** Regulating the supply of money.
   - **d.** Serving as a lender of last resort.
7. James borrows $300,000 for a home from Bank A. Bank A resells the right to collect on that loan to Bank B. Bank B securitizes that loan with hundreds of others and sells the resulting security to a state pension plan, which at the same time purchases an insurance policy from AIG that will pay off if James and the other people whose mortgages are in the security can't pay off their mortgage loans. Suppose that James and all the other people can't pay off their mortgages. Which financial entity is legally obligated to suffer the loss? **LO32.6**
   - **a.** Bank A.
   - **b.** Bank B.
   - **c.** The state pension plan.
   - **d.** AIG.
8. City Bank is considering making a $50 million loan to a company named SheetOil that wants to commercialize a process for turning used blankets, pillowcases, and sheets into oil. This company's chances for success are dubious, but City Bank makes the loan anyway because it believes that the government will bail it out if SheetOil goes bankrupt and cannot repay the loan. City Bank's decision to make the loan has been affected by: **LO32.7**
   - **a.** Liquidity.
   - **b.** Moral hazard.
   - **c.** Token money.
   - **d.** Securitization.
9. True or False: The financial crisis hastened the ongoing process in which the financial services industry was transforming from having a few large firms to many small firms. **LO32.8**

PROBLEMS

1. Assume that the following asset values (in millions of dollars) exist in Ironmania: Federal Reserve Notes in circulation = $700; Money market mutual funds (MMMFs) held by individuals = $400; Corporate bonds = $300; Iron ore deposits = $50; Currency in commercial banks = $100; Savings deposits, including money market deposit accounts

(MMDAs) = $140; Checkable deposits = $1,500; Small-denominated (less than $100,000) time deposits = $100; Coins in circulation = $40. **LO32.1**

a. What is *M*1 in Ironmania?

b. What is *M*2 in Ironmania?

2. Assume that Jimmy Cash has $2,000 in his checking account at Folsom Bank and uses his checking account card to withdraw $200 of cash from the bank's ATM machine. By what dollar amount did the *M*1 money supply change as a result of this single, isolated transaction? **LO32.2**

3. Suppose the price level and value of the U.S. dollar in year 1 are 1 and $1, respectively. If the price level rises to 1.25 in year 2, what is the new value of the dollar? If, instead, the price level falls to 0.50, what is the value of the dollar? **LO32.3**

4. Assume that securitization combined with borrowing and irrational exuberance in Hyperville have driven up the value of existing financial securities at a geometric rate, specifically from $2 to $4 to $8 to $16 to $32 to $64 over a six-year time period. Over the same period, the value of the assets underlying the securities rose at an arithmetic rate from $2 to $3 to $4 to $5 to $6 to $7. If these patterns hold for decreases as well as for increases, by how much would the value of the financial securities decline if the value of the underlying asset suddenly and unexpectedly fell by $5? **LO32.6**

5. Suppose that Lady Gaga goes to Las Vegas to play poker and at the last minute her record company says it will reimburse her for 50 percent of any gambling losses that she incurs. Will Lady Gaga wager more or less as a result of the reimbursement offer? What economic concept does your answer illustrate? **LO32.7**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **33**

Money Creation

Learning Objectives

LO33.1 **Discuss why the U.S. banking system is called a "fractional reserve" system.**

LO33.2 **Explain the basics of a bank's balance sheet and the distinction between a bank's actual reserves and its required reserves.**

LO33.3 **Describe how a bank can create money.**

LO33.4 **Describe the multiple expansion of loans and money by the entire banking system.**

LO33.5 **Define the monetary multiplier, explain how to calculate it, and demonstrate its relevance.**

We have seen that the $M1$ money supply consists of currency (coins and Federal Reserve Notes) and checkable deposits and that $M1$ is a base component of $M2$, a broader measure of the money supply that also includes savings deposits, small-denominated time deposits, and balances in money market mutual funds. The U.S. Mint produces the coins and the U.S. Bureau of Engraving and Printing creates the Federal Reserve Notes. So who creates the checkable deposits? Surprisingly, it is loan officers! Although that may sound like something a congressional committee should investigate, the monetary authorities are well aware that banks and thrifts create checkable deposits. In fact, the Federal Reserve relies on these institutions to create this vital component of the nation's money supply.

The Fractional Reserve System

LO33.1 Discuss why the U.S. banking system is called a "fractional reserve" system.

The United States, like most other countries today, has a **fractional reserve banking system** in which only a portion (fraction) of checkable deposits are backed up by reserves of currency in bank vaults or deposits at the central bank. Our goal is to explain this system and show how commercial banks can create checkable deposits by issuing loans. Our examples will involve commercial banks, but remember that thrift institutions also provide checkable deposits. So the analysis applies to banks and thrifts alike.

Illustrating the Idea: The Goldsmiths

Here is the history behind the idea of the fractional reserve system.

When early traders began to use gold in making transactions, they soon realized that it was both unsafe and inconvenient to carry gold and to have it weighed and assayed (judged for purity) every time they negotiated a transaction. So by the sixteenth century they had begun to deposit their gold with goldsmiths, who would store it in vaults for a fee. On receiving a gold deposit, the goldsmith would issue a receipt to the depositor. Soon people were paying for goods with goldsmiths' receipts, which served as one of the first types of paper money.

At this point the goldsmiths—embryonic bankers—used a 100 percent reserve system; they backed their circulating paper money receipts fully with the gold that they held "in reserve" in their vaults. But because of the public's acceptance of the goldsmiths' receipts as paper money, the goldsmiths soon realized that owners rarely redeemed the gold they had in storage. In fact, the goldsmiths observed that the amount of gold being deposited with them in any week or month was likely to exceed the amount that was being withdrawn.

Then some clever goldsmith hit on the idea that paper "receipts" could be issued in excess of the amount of gold held. Goldsmiths would put these receipts, which were redeemable in gold, into circulation by making interest-earning loans to merchants, producers, and consumers. A borrower might, for instance, borrow $10,000 worth of gold receipts today with the promise to repay $10,500 worth of gold receipts in one year (a 5 percent interest rate). Borrowers were willing to accept loans in the form of gold receipts because the receipts were accepted as a medium of exchange in the marketplace.

This was the beginning of the fractional reserve system of banking, in which reserves in bank vaults are a fraction of the total money supply. If, for example, the goldsmith issued $1 million in receipts for actual gold in storage and another $1 million in receipts as loans, then the total value of paper money in circulation would be $2 million—twice the value of the gold. Gold reserves would be a fraction (one-half) of outstanding paper money.

Significant Characteristics of Fractional Reserve Banking

The goldsmith story highlights two significant characteristics of fractional reserve banking. First, banks can create money through lending. In fact, goldsmiths created money when they made loans by giving borrowers paper money that was not fully backed by gold reserves. The quantity of such money goldsmiths could create depended on the amount of reserves they deemed prudent to have available. The smaller the amount of reserves thought necessary, the larger the amount of paper money the goldsmiths could create. Today, gold is no longer used as bank reserves. Instead, currency itself serves as bank reserves so that the creation of checkable-deposit money by banks (via their lending) is limited by the amount of *currency reserves* that the banks feel obligated, or are required by law, to keep.

A second reality is that banks operating on the basis of fractional reserves are vulnerable to "panics" or "runs." A goldsmith who issued paper money equal to twice the value of his gold reserves would be unable to convert all that paper money into gold in the event that all the holders of that money appeared at his door at the same time demanding their gold. In fact, many European and U.S. banks were once ruined by this unfortunate circumstance. However, a bank panic is highly unlikely if the banker's reserve and lending policies are prudent. Indeed, one reason why banking systems are highly regulated industries is to prevent runs on banks.

This is also why the United States has the system of deposit insurance that we discussed in the last chapter. By guaranteeing deposits, deposit insurance helps to prevent the sort of bank runs that used to happen so often before deposit insurance was available. In those situations, rumors would spread that a bank was about to go bankrupt and that it only had a small amount of reserves left in its vaults. Bank runs are called "bank runs" because depositors would run to the bank trying to be one of the lucky few to withdraw their money while the bank had any reserves left. The rumors were usually totally unfounded. But, unfortunately, the bank would still go bankrupt even if it began the day with its normal amount of reserves. With so many customers withdrawing money simultaneously, it would run out of reserves and be forced to default on its obligations to its remaining depositors. By guaranteeing depositors that they will always get their money, deposit insurance removes the incentive to try to withdraw one's deposit before anyone else can. It thus stops most bank runs.

A Single Commercial Bank

LO33.2 Explain the basics of a bank's balance sheet and the distinction between a bank's actual reserves and its required reserves.

To illustrate the workings of the modern fractional reserve banking system, we need to examine a commercial bank's balance sheet.

The **balance sheet** of a commercial bank (or thrift) is a statement of assets—things owned by the bank or owed to the bank—and claims on those assets. A bank balance sheet summarizes the financial position of the bank at a certain time. Every balance sheet must balance; this means that the value of *assets* must equal the amount of claims against those assets. The claims shown on a balance sheet are divided into two groups: the claims of nonowners of the bank against the firm's assets, called *liabilities*, and the claims of the owners of the firm against the firm's assets, called *net worth*. Liabilities are things owed by the bank to depositors or others. A balance sheet is balanced because

$$\text{Assets} = \text{liabilities} + \text{net worth}$$

Every \$1 change in assets must be offset by a \$1 change in liabilities + net worth. Every \$1 change in liabilities + net worth must be offset by a \$1 change in assets.

Now let's work through a series of bank transactions involving balance sheets to establish how individual banks can create money.

Transaction 1: Creating a Bank

Suppose some far-sighted citizens of the town of Wahoo, Nebraska (yes, there is such a place), decide their town needs a new commercial bank to provide banking services for that growing community. Once they have secured a state or national charter for their bank, they turn to the task of selling, say, \$250,000 worth of stock (equity shares) to buyers, both in and out of the community. Their efforts meet with success and the Bank of Wahoo comes into existence—at least on paper. What does its balance sheet look like at this stage?

The founders of the bank have sold \$250,000 worth of shares of stock in the bank—some to themselves, some to other people. As a result, the bank now has \$250,000 in cash on hand and \$250,000 worth of stock shares outstanding. The cash is an asset to the bank. Cash held by a bank is sometimes called **vault cash** or till money. The shares of stock outstanding constitute an equal amount of claims that the owners have against the bank's assets. Those shares of stock constitute the net worth of the bank. The bank's balance sheet reads:

Creating a Bank
Balance Sheet 1: Wahoo Bank

Assets		Liabilities and net worth	
Cash	\$250,000	Stock shares	\$250,000

Each item listed in a balance sheet such as this is called an *account*.

Transaction 2: Acquiring Property and Equipment

The board of directors (who represent the bank's owners) must now get the new bank off the drawing board and make it a reality. First, property and equipment must be acquired. Suppose the directors, confident of the success of their venture, purchase a building for \$220,000 and pay \$20,000 for office equipment. This simple transaction changes the composition of the bank's assets. The bank now has \$240,000 less in cash and \$240,000 of new property assets. Using blue to denote accounts affected by each transaction, we find that the bank's balance sheet at the end of transaction 2 appears as follows:

Acquiring Property and Equipment
Balance Sheet 2: Wahoo Bank

Assets		Liabilities and net worth	
Cash	\$ 10,000	Stock shares	\$250,000
Property	240,000		

Note that the balance sheet still balances, as it must.

Transaction 3: Accepting Deposits

Commercial banks have two basic functions: to accept deposits of money and to make loans. Now that the bank is operating, suppose that the citizens and businesses of Wahoo decide to deposit \$100,000 in the Wahoo bank. What happens to the bank's balance sheet?

The bank receives cash, which is an asset to the bank. Suppose this money is deposited in the bank as checkable deposits (checking account entries), rather than as savings accounts or time deposits. These newly created *checkable deposits* constitute claims that the depositors have against the assets of the Wahoo bank and thus are a new liability account. The bank's balance sheet now looks like this:

Accepting Deposits Balance Sheet 3: Wahoo Bank			
Assets		**Liabilities and net worth**	
Cash	$110,000	Checkable deposits	$100,000
Property	240,000	Stock shares	250,000

There has been no change in the economy's total supply of money as a result of transaction 3, but a change has occurred in the composition of the money supply. Bank money, or checkable deposits, has increased by $100,000, and currency held by the public has decreased by $100,000. As explained in the previous chapter, currency held in a bank is not part of the economy's money supply.

A withdrawal of cash will reduce the bank's checkable-deposit liabilities and its holdings of cash by the amount of the withdrawal. This, too, changes the composition, but not the total supply, of money in the economy.

Transaction 4: Depositing Reserves in a Federal Reserve Bank

All commercial banks and thrift institutions that provide checkable deposits must by law keep **required reserves.** Required reserves are an amount of funds equal to a specified percentage of the bank's own deposit liabilities. A bank must keep these reserves on deposit with the Federal Reserve Bank in its district or as cash in the bank's vault. To simplify, we suppose the Bank of Wahoo keeps its required reserves entirely as deposits in the Federal Reserve Bank of its district. But remember that vault cash is counted as reserves and real-world banks keep a significant portion of their own reserves in their vaults.

The "specified percentage" of checkable-deposit liabilities that a commercial bank must keep as reserves is known as the **reserve ratio**—the ratio of the required reserves the commercial bank must keep to the bank's own outstanding checkable-deposit liabilities:

$$\text{Reserve ratio} = \frac{\text{commercial bank's required reserves}}{\text{commercial bank's checkable-deposit liabilities}}$$

If the reserve ratio is $\frac{1}{10}$, or 10 percent, the Wahoo bank, having accepted $100,000 in deposits from the public, would have to keep $10,000 as reserves. If the ratio is $\frac{1}{5}$, or 20 percent, $20,000 of reserves would be required. If $\frac{1}{2}$, or 50 percent, $50,000 would be required.

TABLE 33.1 Reserve Requirements (Reserve Ratios) for Banks and Thrifts, 2013

Type of Deposit	Current Requirement	Statutory Limits
Checkable deposits:		
$0–$12.4 million	0%	3%
$12.4–$79.5 million	3	3
Over $79.5 million	10	8–14
Noncheckable nonpersonal savings and time deposits	0	0–9

Source: Federal Reserve, Regulation D, **www.federalreserve.gov**.

The Fed has the authority to establish and vary the reserve ratio within limits legislated by Congress. The limits now prevailing are shown in Table 33.1. The first $12.4 million of checkable deposits held by a commercial bank or thrift is exempt from reserve requirements. A 3 percent reserve is required on checkable deposits of between $12.4 million and $79.5 million. A 10 percent reserve is required on checkable deposits over $79.5 million, although the Fed can vary that percentage between 8 and 14 percent. Currently, no reserves are required against noncheckable nonpersonal (business) savings or time deposits, although up to 9 percent can be required. Also, after consultation with appropriate congressional committees, the Fed for 180 days may impose reserve requirements outside the 8 to 14 percent range specified in Table 33.1. Beginning in late 2008, the Fed began paying banks interest on their required reserves and on their excess reserve balances held at Federal Reserve Banks.

In order to simplify, we will suppose that the reserve ratio for checkable deposits in commercial banks is $\frac{1}{5}$, or 20 percent. Although 20 percent obviously is higher than the requirement really is, the figure is convenient for calculations. Because we are concerned only with checkable (spendable) deposits, we ignore reserves on noncheckable savings and time deposits. The main point is that reserve requirements are fractional, meaning that they are less than 100 percent. This point is critical in our analysis of the lending ability of the banking system.

By depositing $20,000 in the Federal Reserve Bank, the Wahoo bank will just be meeting the required 20 percent ratio between its reserves and its own deposit liabilities. We will use "reserves" to mean the funds commercial banks deposit in the Federal Reserve Banks, to distinguish those funds from the public's deposits in commercial banks.

But suppose the Wahoo bank anticipates that its holdings of checkable deposits will grow in the future. Then, instead of sending just the minimum amount, $20,000, it sends an extra $90,000, for a total of $110,000. In so doing, the bank will avoid the inconvenience of sending additional reserves to the Federal Reserve Bank each time its own checkable-deposit liabilities increase. And, as you will see, it is these extra reserves that enable banks to lend money and earn interest income.

Actually, a real-world bank would not deposit *all* its cash in the Federal Reserve Bank. However, because (1) banks as a rule hold vault cash only in the amount of $1\frac{1}{2}$ or 2 percent of their total assets and (2) vault cash can be counted as reserves, we will assume for simplicity that all of Wahoo's cash is deposited in the Federal Reserve Bank and therefore constitutes the commercial bank's actual reserves. By making this simplifying assumption, we do not need to bother adding two assets—"cash" and "deposits in the Federal Reserve Bank"—to determine "reserves."

After the Wahoo bank deposits $110,000 of reserves at the Fed, its balance sheet becomes:

Depositing Reserves at the Fed
Balance Sheet 4: Wahoo Bank

Assets		Liabilities and net worth	
Cash	$ 0	Checkable deposits	$100,000
Reserves	110,000	Stock shares	250,000
Property	240,000		

There are three things to note about this latest transaction.

Excess Reserves A bank's **excess reserves** are found by subtracting its *required reserves* from its **actual reserves:**

$$\text{Excess reserves} = \text{actual reserves} - \text{required reserves}$$

In this case,

Actual reserves	$110,000
Required reserves	−20,000
Excess reserves	$ 90,000

The only reliable way of computing excess reserves is to multiply the bank's checkable-deposit liabilities by the reserve ratio to obtain required reserves ($100,000 × 20 percent = $20,000) and then to subtract the required reserves from the actual reserves listed on the asset side of the bank's balance sheet.

To test your understanding, compute the bank's excess reserves from balance sheet 4, assuming that the reserve ratio is (1) 10 percent, (2) $33\frac{1}{3}$ percent, and (3) 50 percent.

We will soon demonstrate that the ability of a commercial bank to make loans depends on the existence of excess reserves. Understanding this concept is crucial in seeing how the banking system creates money.

Control You might think the basic purpose of reserves is to enhance the liquidity of a bank and protect commercial bank depositors from losses. Reserves would constitute a ready source of funds from which commercial banks could meet large, unexpected cash withdrawals by depositors.

But this reasoning breaks down under scrutiny. Although historically reserves have been seen as a source of liquidity and therefore as protection for depositors, a bank's required reserves are not great enough to meet sudden, massive cash withdrawals. If the banker's nightmare should materialize—everyone with checkable deposits appearing at once to demand those deposits in cash—the actual reserves held as vault cash or at the Federal Reserve Bank would be insufficient. The banker simply could not meet this "bank panic." Because reserves are fractional, checkable deposits may be much greater than a bank's required reserves.

So commercial bank deposits must be protected by other means. Periodic bank examinations are one way of promoting prudent commercial banking practices. Furthermore, insurance funds administered by the Federal Deposit Insurance Corporation (FDIC) and the National Credit Union Administration (NCUA) insure individual deposits in banks and thrifts up to $250,000.

If it is not the purpose of reserves to provide for commercial bank liquidity, then what is their function? *Control* is the answer. Required reserves help the Fed control the lending ability of commercial banks. The Fed can take certain actions that either increase or decrease commercial bank reserves and affect the ability of banks to grant credit. The objective is to prevent banks from overextending or underextending bank credit. To the degree that these policies successfully influence the volume of commercial bank credit, the Fed can help the economy avoid business fluctuations. Another function of reserves is to facilitate the collection or "clearing" of checks.

Asset and Liability Transaction 4 brings up another matter. Specifically, the reserves created in transaction 4 are an asset to the depositing commercial bank because they are a claim this bank has against the assets of another institution—the Federal Reserve Bank. The checkable deposit you get by depositing money in a commercial bank is an asset to you and a liability to the bank (since the bank is liable for repaying you whenever you choose to withdraw your deposit). In the same way, the reserves that a commercial bank establishes by depositing money in a bankers' bank are an asset to the commercial bank and a liability to the Federal Reserve Bank.

Transaction 5: Clearing a Check Drawn against the Bank

Assume that Fred Bradshaw, a Wahoo farmer, deposited a substantial portion of the $100,000 in checkable deposits that the Wahoo bank received in transaction 3. Now suppose that Fred buys $50,000 of farm machinery from the Ajax Farm Implement Company of Surprise, Nebraska. Bradshaw pays for this machinery by writing a $50,000 check against his deposit in the Wahoo bank. He gives the check to the Ajax Company. What are the results?

Ajax deposits the check in its account with the Surprise bank. The Surprise bank increases Ajax's checkable deposits by $50,000 when Ajax deposits the check. Ajax is now paid in full. Bradshaw is pleased with his new machinery.

Now the Surprise bank has Bradshaw's check. This check is simply a claim against the assets of the Wahoo bank. The Surprise bank will collect this claim by sending the check (along with checks drawn on other banks) to the regional Federal Reserve Bank. Here a bank employee will clear, or collect, the check for the Surprise bank by increasing Surprise's reserve in the Federal Reserve Bank by $50,000 and decreasing the Wahoo bank's reserve by that same amount. The check is "collected" merely by making bookkeeping notations to the effect that Wahoo's claim against the Federal Reserve Bank is reduced by $50,000 and Surprise's claim is increased by $50,000.

Finally, the Federal Reserve Bank sends the cleared check back to the Wahoo bank, and for the first time the Wahoo bank discovers that one of its depositors has drawn a check for $50,000 against his checkable deposit. Accordingly, the Wahoo bank reduces Bradshaw's checkable deposit by $50,000 and notes that the collection of this check has caused a $50,000 decline in its reserves at the Federal Reserve Bank. All the balance sheets balance: The Wahoo bank has reduced both its assets (reserves) and its liabilities (checkable deposits) by $50,000. The Surprise bank has $50,000 more in both assets (reserves) and liabilities (checkable deposits). Ownership of reserves at the Federal Reserve Bank has changed—with Wahoo owning $50,000 less and Surprise owning $50,000 more—but total reserves stay the same.

Whenever a check is drawn against one bank and deposited in another bank, collection of that check will reduce both the reserves and the checkable deposits of the bank on which the check is drawn. Conversely, if a bank receives a check drawn on another bank, the bank receiving the check will, in the process of collecting it, have its reserves and deposits increased by the amount of the check. In our example, the Wahoo bank loses $50,000 in both reserves and deposits to the Surprise bank. But there is no loss of reserves or deposits for the banking system as a whole. What one bank loses, another bank gains.

If we bring all the other assets and liabilities back into the picture, the Wahoo bank's balance sheet looks like this at the end of transaction 5:

Clearing a Check
Balance Sheet 5: Wahoo Bank

Assets		Liabilities and net worth	
Reserves	$ 60,000	Checkable deposits	$ 50,000
Property	240,000	Stock shares	250,000

Verify that with a 20 percent reserve requirement, the bank's excess reserves now stand at $50,000.

QUICK REVIEW 33.1

- The United States has a fractional reserve banking system, in which the collective reserves of the banks usually are considerably less than 100 percent of their checkable deposit liabilities.
- When a bank accepts deposits of cash, the composition of the money supply is changed, but the total supply of money is not directly altered.
- Commercial banks and thrifts are obliged to keep required reserves equal to a specified percentage of their own checkable-deposit liabilities as cash or on deposit with the Federal Reserve Bank of their district.
- The amount by which a bank's actual reserves exceed its required reserves is called excess reserves.
- A bank that has a check drawn and collected against it will lose to the recipient bank both reserves and deposits equal to the value of the check.

Money-Creating Transactions of a Commercial Bank

LO33.3 Describe how a bank can create money.

The next two transactions are crucial because they explain (1) how a commercial bank can literally create money by making loans and (2) how banks create money by purchasing government bonds from the public.

Transaction 6: Granting a Loan

In addition to accepting deposits, commercial banks grant loans to borrowers. What effect does lending by a commercial bank have on its balance sheet?

Suppose the Gristly Meat Packing Company of Wahoo decides it is time to expand its facilities. Suppose, too, that the company needs exactly $50,000—which just happens to be equal to the Wahoo bank's excess reserves—to finance this project.

Gristly goes to the Wahoo bank and requests a loan for this amount. The Wahoo bank knows the Gristly Company's fine reputation and financial soundness and is convinced of its ability to repay the loan. So the loan is granted. In return, the president of Gristly hands a promissory note—a fancy IOU—to the Wahoo bank. Gristly wants the convenience and safety of paying its obligations by check. So, instead of receiving a basket full of currency from the bank, Gristly gets a $50,000 increase in its checkable-deposit account in the Wahoo bank.

The Wahoo bank has acquired an interest-earning asset (the promissory note, which it files under "Loans") and has created checkable deposits (a liability) to "pay" for this asset. Gristly has swapped an IOU for the right to draw an additional $50,000 worth of checks against its checkable deposit in the Wahoo bank. Both parties are pleased.

At the moment the loan is completed, the Wahoo bank's position is shown by balance sheet 6a:

When a Loan Is Negotiated
Balance Sheet 6a: Wahoo Bank

Assets		Liabilities and net worth	
Reserves	$ 60,000	Checkable deposits	$100,000
Loans	50,000	Stock shares	250,000
Property	240,000		

All this looks simple enough. But a close examination of the Wahoo bank's balance statement reveals a startling fact: When a bank makes loans, it creates money. The president of Gristly went to the bank with something that is *not* money—her IOU—and walked out with something that *is* money—a checkable deposit.

Contrast transaction 6a with transaction 3, in which checkable deposits were created but only as a result of currency having been taken out of circulation. There was a change in the *composition* of the money supply in that situation but no change in the *total supply* of money. But when banks lend, they create checkable deposits that *are* money. By extending credit, the Wahoo bank has "monetized" an IOU. Gristly and the Wahoo bank have created and then swapped claims. The claim created by Gristly and given to the bank is not money; an individual's IOU is not acceptable as a medium of exchange. But the claim created by the bank and given to Gristly *is* money; checks drawn against a checkable deposit are acceptable as a medium of exchange. Checkable-deposit money like this constitutes about one-half the quantity of *M*1 money in the United States and about 10 percent of *M*2.

Much of the money used in the United States therefore is created through the extension of credit by commercial banks. This checkable-deposit money may be thought of as "debts" of commercial banks and thrift institutions. Checkable deposits are bank debts in the sense that they are claims that banks and thrifts promise to pay "on demand."

But certain factors limit the ability of a commercial bank to create checkable deposits ("bank money") by lending. The Wahoo bank can expect the newly created checkable deposit of $50,000 to be a very active account. Gristly would not borrow $50,000 at, say, 7, 10, or 12 percent interest for the sheer joy of knowing that funds were available if needed.

Assume that Gristly awards a $50,000 building contract to the Quickbuck Construction Company of Omaha. Quickbuck, true to its name, completes the expansion promptly and is paid with a check for $50,000 drawn by Gristly against its checkable deposit in the Wahoo bank. Quickbuck, with headquarters in Omaha, does not deposit this check in the Wahoo bank but instead deposits it in the Fourth National Bank of Omaha. Fourth National now has a $50,000 claim against the Wahoo bank. The check is collected in the manner described in transaction 5. As a result, the Wahoo bank loses both reserves and deposits equal to the amount of the check; Fourth National acquires $50,000 of reserves and deposits.

In summary, assuming a check is drawn by the borrower for the entire amount of the loan ($50,000) and is given to a firm that deposits it in some other bank, the Wahoo bank's balance sheet will read as follows after the check has been cleared against it:

After a Check Is Drawn on the Loan
Balance Sheet 6b: Wahoo Bank

Assets		Liabilities and net worth	
Reserves	$ 10,000	Checkable deposits	$ 50,000
Loans	50,000	Stock shares	250,000
Property	240,000		

After the check has been collected, the Wahoo bank just meets the required reserve ratio of 20 percent (= $10,000/$50,000). The bank has *no* excess reserves. This poses a question: Could the Wahoo bank have lent more than $50,000—an amount greater than its excess reserves—and still have met the 20 percent reserve requirement when a check for the full amount of the loan was cleared against it? The answer is no; the bank is "fully loaned up."

Here is why: Suppose the Wahoo bank had lent $55,000 to the Gristly company and that the Gristly company had spent all of that money by writing a $55,000 check to

WORKED PROBLEMS
W33.1
Single bank accounting

Quickbuck Construction. Collection of the check against the Wahoo bank would have lowered its reserves to $5,000 (= $60,000 − $55,000), and checkable deposits would once again stand at $50,000 (= $105,000 − $55,000). The ratio of actual reserves to checkable deposits would then be $5,000/$50,000, or only 10 percent. Because the reserve requirement is 20 percent, the Wahoo bank could not have lent $55,000.

By experimenting with other amounts over $50,000, you will find that the maximum amount the Wahoo bank could lend at the outset of transaction 6 is $50,000. This amount is identical to the amount of excess reserves the bank had available when the loan was negotiated.

A single commercial bank in a multibank banking system can lend only an amount equal to its initial preloan excess reserves. When it lends, the lending bank faces the possibility that checks for the entire amount of the loan will be drawn and cleared against it. If that happens, it will lose (to other banks) reserves equal to the amount it lends. So, to be safe, it limits its lending to the amount of its excess reserves.

Bank creation of money raises an interesting question: If a bank creates checkable-deposit money when it lends its excess reserves, is money destroyed when borrowers pay off loans? The answer is yes. When loans are paid off, the process works in reverse. The bank's checkable deposits decline by the amount of the loan repayment.

Transaction 7: Buying Government Securities

When a commercial bank buys government bonds from the public, the effect is substantially the same as lending. New money is created.

Assume that the Wahoo bank's balance sheet initially stands as it did at the end of transaction 5. Now suppose that instead of making a $50,000 loan, the bank buys $50,000 of government securities from a securities dealer. The bank receives the interest-bearing bonds, which appear on its balance statement as the asset "Securities," and gives the dealer an increase in its checkable-deposit account. The Wahoo bank's balance sheet appears as follows:

Buying Government Securities
Balance Sheet 7: Wahoo Bank

Assets		Liabilities and net worth	
Reserves	$ 60,000	Checkable deposits	$100,000
Securities	50,000	Stock shares	250,000
Property	240,000		

Checkable deposits, that is, the supply of money, have been increased by $50,000, as in transaction 6. Bond purchases from the public by commercial banks increase the supply of money in the same way as lending to the public does. The bank accepts government bonds (which are not money) and gives the securities dealer an increase in its checkable deposits (which *are* money).

Of course, when the securities dealer draws and clears a check for $50,000 against the Wahoo bank, the bank loses both reserves and deposits in that amount and then just meets the legal reserve requirement. Its balance sheet now reads precisely as in 6b except that "Securities" is substituted for "Loans" on the asset side.

Finally, the *selling* of government bonds to the public by a commercial bank—like the repayment of a loan—reduces the supply of money. The securities buyer pays by check, and both "Securities" and "Checkable deposits" (the latter being money) decline by the amount of the sale.

Profits, Liquidity, and the Federal Funds Market

The asset items on a commercial bank's balance sheet reflect the banker's pursuit of two conflicting goals:

- ***Profit*** One goal is profit. Commercial banks, like any other businesses, seek profits, which is why the bank makes loans and buys securities—the two major earning assets of commercial banks.
- ***Liquidity*** The other goal is safety. For a bank, safety lies in liquidity, specifically such liquid assets as cash and excess reserves. A bank must be on guard for depositors who want to transform their checkable deposits into cash. Similarly, it must guard against more checks clearing against it than are cleared in its favor, causing a net outflow of reserves. Bankers thus seek a balance between prudence and profit. The compromise is between assets that earn higher returns and highly liquid assets that earn no returns.

An interesting way in which banks can partly reconcile the goals of profit and liquidity is to lend temporary excess reserves held at the Federal Reserve Banks to other commercial banks. Normal day-to-day flows of funds to banks rarely leave all banks with their exact levels of required reserves. Also, excess reserves held at the Federal Reserve Banks are highly liquid, but they draw less interest than the banks can make through loans. Banks therefore often lend these excess reserves to other banks on an overnight basis in order to earn additional interest without sacrificing long-term liquidity. Banks that borrow in this federal funds market—the market for immediately available reserve balances at the Federal Reserve—do so because they are temporarily short of required reserves. The interest rate paid on these overnight loans is called the **federal funds rate.**

We would show an overnight loan of reserves from the Surprise bank to the Wahoo bank as a decrease in reserves at the Surprise bank and an increase in reserves at the Wahoo bank. Ownership of reserves at the Federal Reserve Bank of Kansas City would change, but total reserves would not be affected. Exercise: Determine what other changes would be required on the Wahoo and Surprise banks' balance sheets as a result of the overnight loan.

QUICK REVIEW 33.2

- Banks create money when they make loans; money vanishes when bank loans are repaid.
- New money is created when banks buy government bonds from the public; money disappears when banks sell government bonds to the public.
- Banks balance profitability and safety in determining their mix of earning assets and highly liquid assets.
- Although the Fed pays interest on excess reserves, banks may be able to obtain higher interest rates by temporarily lending the reserves to other banks in the federal funds market; the interest rate on such loans is the federal funds rate.

The Banking System: Multiple-Deposit Expansion

LO33.4 Describe the multiple expansion of loans and money by the entire banking system.

Thus far we have seen that a single bank in a banking system can lend one dollar for each dollar of its excess reserves. The situation is different for all commercial banks as a group. We will find that the commercial banking system can lend—that is, can create money—by a multiple of its excess reserves. This multiple lending is accomplished even though each bank in the system can lend only "dollar for dollar" with its excess reserves.

How do these seemingly paradoxical results come about? To answer this question succinctly, we will make three simplifying assumptions:

- The reserve ratio for all commercial banks is 20 percent.
- Initially all banks are meeting this 20 percent reserve requirement exactly. No excess reserves exist; or, in the parlance of banking, they are "loaned up" (or "loaned out") fully in terms of the reserve requirement.
- If any bank can increase its loans as a result of acquiring excess reserves, an amount equal to those excess reserves will be lent to one borrower, who will write a check for the entire amount of the loan and give it to someone else, who will deposit the check in another bank. This third assumption means that the worst thing possible happens to every lending bank—a check for the entire amount of the loan is drawn and cleared against it in favor of another bank.

The Banking System's Lending Potential

Suppose a junkyard owner finds a $100 bill while dismantling a car that has been on the lot for years. He deposits the $100 in bank A, which adds the $100 to its reserves. We will record only changes in the balance sheets of the various commercial banks. The deposit changes bank A's balance sheet as shown by entries (a_1):

Multiple-Deposit Expansion Process
Balance Sheet: Commercial Bank A

Assets		Liabilities and net worth	
Reserves	$+100 ($a_1$)	Checkable deposits	$+100 ($a_1$)
	−80 (a_3)		+80 (a_2)
Loans	+80 (a_2)		−80 (a_3)

Recall from transaction 3 that this $100 deposit of currency does not alter the money supply. While $100 of checkable-deposit money comes into being, it is offset by the $100 of currency no longer in the hands of the public (the junkyard owner). But bank A *has* acquired excess reserves of $80. Of the newly acquired $100 in currency, 20 percent, or $20, must be earmarked for the required reserves on the new $100 checkable deposit, and the remaining $80 goes to excess reserves. Remembering that a single commercial bank can lend only an amount equal to its excess reserves, we conclude that bank A can lend a maximum of $80. When a loan for this amount is made, bank A's loans increase by $80 and the borrower gets an $80 checkable deposit. We add these figures—entries (a_2)—to bank A's balance sheet.

But now we make our third assumption: The borrower uses the full amount of the loan ($80) to write a check ($80) to someone else, and that person deposits the amount in bank B, a different bank. As we saw in transaction 6, bank A loses both reserves and deposits equal to the amount of the loan, as indicated in entries (a_3). The net result of these transactions is that bank A's reserves now stand at +$20 (= $100 − $80), loans at +$80, and checkable deposits at +$100 (= $100 + $80 − $80).

When the dust has settled, bank A is just meeting the 20 percent reserve ratio.

Recalling our previous discussion, we know that bank B acquires both the reserves and the deposits that bank A has lost. Bank B's balance sheet is changed as in entries (b_1):

Multiple-Deposit Expansion Process
Balance Sheet: Commercial Bank B

Assets		Liabilities and net worth	
Reserves	$+80 ($b_1$)	Checkable deposits	$+80 ($b_2$)...

When the borrower's check is drawn and cleared, bank A loses $80 in reserves and deposits and bank B gains $80 in reserves and deposits. But 20 percent, or $16, of bank B's new reserves must be kept as required reserves against the new $80 in checkable deposits. This means that bank B has $64 (= $80 − $16) in excess reserves. It can therefore lend $64 [entries ($b_2$)]. When the new borrower writes a check for $64 to buy a product, and the seller deposits the check in bank C, the reserves and deposits of bank B both fall by $64 [entries ($b_3$)]. As a result of these transactions, bank B's reserves now stand at +$16 (= $80 − $64), loans at +$64, and checkable deposits at +$80 (= $80 + $64 − $64). After all this, bank B is just meeting the 20 percent reserve requirement.

We are off and running again. Bank C acquires the $64 in reserves and deposits lost by bank B. Its balance sheet changes as in entries (c_1):

Multiple-Deposit Expansion Process
Balance Sheet: Commercial Bank C

Assets		Liabilities and net worth	
Reserves	$+64.00 ($c_1$)	Checkable deposits	$+64.00 ($c_1$)
	−51.20 (c_3)		+51.20 (c_2)
Loans	+51.20 (c_2)		−51.20 (c_3)

Exactly 20 percent, or $12.80, of these new reserves will be required reserves, the remaining $51.20 being excess reserves. Hence, bank C can safely lend a maximum of $51.20. Suppose it does [entries (c_2)]. And suppose the borrower writes a check for the entire amount ($51.20) to a merchant who deposits it in another bank [entries (c_3)].

We could go ahead with this procedure by bringing banks D, E, F, G, . . . , N, and so on into the picture. In fact, the process will go on almost indefinitely, just as long as banks further down the line receive at least one penny in new reserves that they can use to back another round of lending and money creation. But we suggest that you work through the computations for banks D, E, and F to be sure you understand the procedure.

The entire analysis is summarized in Table 33.2. Data for banks D through N are supplied on their own rows so

TABLE 33.2 Expansion of the Money Supply by the Commercial Banking System

Bank	(1) Acquired Reserves and Deposits	(2) Required Reserves (Reserve Ratio = .2)	(3) Excess Reserves, (1) − (2)	(4) Amount Bank Can Lend; New Money Created = (3)
Bank A	$100.00 ($a_1$)	$20.00	$80.00	$ 80.00 (a_2)
Bank B	80.00 (a_3, b_1)	16.00	64.00	64.00 (b_2)
Bank C	64.00 (b_3, c_1)	12.80	51.20	51.20 (c_2)
Bank D	51.20	10.24	40.96	40.96
Bank E	40.96	8.19	32.77	32.77
Bank F	32.77	6.55	26.21	26.21
Bank G	26.21	5.24	20.97	20.97
Bank H	20.97	4.20	16.78	16.78
Bank I	16.78	3.36	13.42	13.42
Bank J	13.42	2.68	10.74	10.74
Bank K	10.74	2.15	8.59	8.59
Bank L	8.59	1.72	6.87	6.87
Bank M	6.87	1.37	5.50	5.50
Bank N	5.50	1.10	4.40	4.40
Other banks	21.99	4.40	17.59	17.59
Total amount of money created (sum of the amounts in column 4)				$400.00

that you may check your computations. The last row of the table consolidates into one row everything that happens for all banks down the line after bank N. Our conclusion is startling: On the basis of only \$80 in excess reserves (acquired by the banking system when someone deposited \$100 of currency in bank A), the entire commercial banking system is able to lend \$400, the sum of the amounts in column 4. The banking system can lend excess reserves by a multiple of 5 (= \$400/\$80) when the reserve ratio is 20 percent. Yet each single bank in the banking system is lending only an amount equal to its own excess reserves. How do we explain this? How can the banking system as a whole lend by a multiple of its excess reserves, when each individual bank can lend only dollar for dollar with its excess reserves?

The answer is that reserves lost by a single bank are not lost to the banking system as a whole. The reserves lost by bank A are acquired by bank B. Those lost by B are gained by C. C loses to D, D to E, E to F, and so forth. Although reserves can be, and are, lost by individual banks in the banking system, there is no loss of reserves for the banking system as a whole.

An individual bank can safely lend only an amount equal to its excess reserves, *but the commercial banking system can lend by a multiple of its collective excess reserves.* This contrast, incidentally, is an illustration of why it is imperative that we keep the fallacy of composition (Last Word, Chapter 1) firmly in mind. Commercial banks as a group can create money by lending in a manner much different from that of the individual banks in the group.

The Monetary Multiplier

LO33.5 Define the monetary multiplier, explain how to calculate it, and demonstrate its relevance.

The **monetary multiplier** (or, less commonly, the *checkable deposit multiplier*) defines the relationship between any new excess reserves in the banking system and the magnified creation of new checkable-deposit money by banks as a group. It is a separate idea from the spending-income multiplier of Chapter 28 but shares some mathematical similarities. The spending-income multiplier exists because the expenditures of one household become some other household's income; the multiplier magnifies a change in initial spending into a larger change in GDP. The spending-income multiplier is the reciprocal of the MPS (the leakage into saving that occurs at each round of spending).

Similarly, the monetary multiplier exists because the reserves and deposits lost by one bank become reserves of another bank. It magnifies excess reserves into a larger creation of checkable-deposit money. The monetary multiplier m is the reciprocal of the required reserve ratio R (the leakage into required reserves that occurs at each step in the lending process). In short,

$$\text{Monetary multiplier} = \frac{1}{\text{required reserve ratio}}$$

or, in symbols,

$$m = \frac{1}{R}$$

In this formula, m represents the maximum amount of new checkable-deposit money that can be created by a single dollar of excess reserves, given the value of R. By multiplying the excess reserves E by m, we can find the maximum amount of new checkable-deposit money, D, that can be created by the banking system. That is,

$$\begin{array}{l}\text{Maximum} \\ \text{checkable-deposit} \\ \text{creation}\end{array} = \begin{array}{c}\text{excess} \\ \text{reserves}\end{array} \times \begin{array}{c}\text{monetary} \\ \text{multiplier}\end{array}$$

or, more simply,

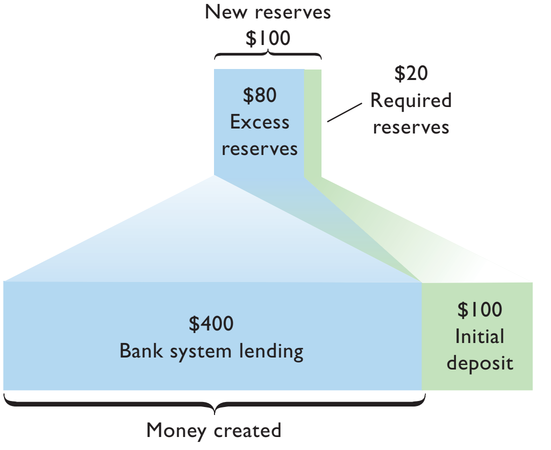
$$D = E \times m$$

In our example in Table 33.2, R is 0.20, so m is 5 (= 1/0.20). This implies that

$$D = \$80 \times 5 = \$400$$

Figure 33.1 depicts the final outcome of our example of a multiple-deposit expansion of the money supply.

FIGURE 33.1 The outcome of the money expansion process. A deposit of \$100 of currency into a checking account creates an initial checkable deposit of \$100. If the reserve ratio is 20 percent, only \$20 of reserves is legally required to support the \$100 checkable deposit. The \$80 of excess reserves allows the banking system to create \$400 of checkable deposits through making loans. The \$100 of reserves supports a total of \$500 of money (\$100 + \$400).

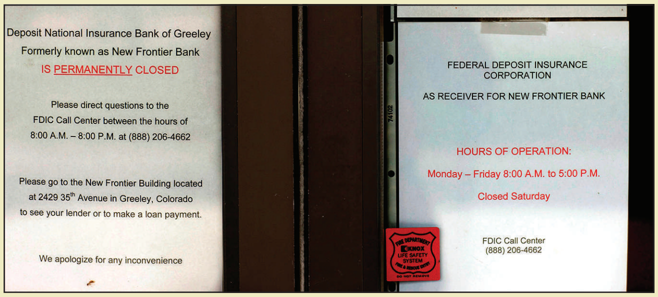


LAST WORD

Banking, Leverage, and Financial Instability

Leverage Boosts Banking Profits but Makes the Banking System Less Stable. Time to Reduce Leverage?

The term *leverage* is used in finance to describe how the use of borrowed money can magnify both profits and losses. To see how leverage works, first consider an investment opportunity that produces a 10 percent positive return if things go well but a 5 percent loss if things go poorly. Those rates of return imply that if a person invests $100 of his own savings, he will end up with either $110 if things go well or $95 if things go badly. Put slightly differently, he will either gain $10 or lose $5 from where he started, with his own $100 being used to fund the $100 investment.



But now consider what happens to his potential returns if he uses borrowed money to provide "leverage." In fact, let's have him use a lot of leverage. To make the $100 investment, he uses $10 of his own savings and $90 of borrowed money (which, for simplicity, we will assume that he can borrow at zero percent interest). If things go well, the investment will return $110. He then must repay the $90 loan. That will leave him with $20 (= $110 of investment return if things go well minus $90 to repay the loan). That means that he will end up with $20 if things go well and he uses leverage. Notice that this implies that if he uses leverage, he will get a 100 percent (= $20 divided by $10) return on the $10 of his own savings that he himself invested. That, of course, is much nicer than a 10 percent return. In fact, it is 10 times as large—hence the term *leverage*. The use of borrowed money has massively increased the percentage rate of return if things go well.

Unfortunately, however, nothing in life is free. Leverage also magnifies the investor's losses if things go wrong. To see this, note that if things go wrong in this example, the investor will end up turning $100 into $95. But then he has to pay back the $90 that he borrowed. That will leave him with only $5 (= $95 investment

The initial deposit of $100 of currency into the bank (lower right-hand box) creates new reserves of an equal amount (upper box). With a 20 percent reserve ratio, however, only $20 of reserves are needed to "back up" this $100 checkable deposit. The excess reserves of $80 permit the creation of $400 of new checkable deposits via the making of loans, confirming a monetary multiplier of 5. The $100 of new reserves supports a total supply of money of $500, consisting of the $100 initial checkable deposit plus $400 of checkable deposits created through lending.

WORKED PROBLEMS

W33.2
Money creation

Higher reserve ratios mean lower monetary multipliers and therefore less creation of new checkable-deposit money via loans; smaller reserve ratios mean higher monetary multipliers and thus more creation of new checkable-deposit money via loans. With a high reserve ratio, say, 50 percent, the monetary multiplier would be 2 (= 1/0.5), and in our example the banking system could create only $100 (= $50 of excess reserves × 2) of new checkable deposits. With a low reserve ratio, say, 5 percent, the monetary multiplier would be 20 (= 1/0.05), and the banking system could create $1,900 (= $95 of excess reserves × 20) of new checkable deposits.

You might experiment with the following two brainteasers to test your understanding of multiple credit expansion by the banking system:

- Rework the analysis in Table 33.2 (at least three or four steps of it) assuming the reserve ratio is 10 percent. What is the maximum amount of money the banking system can create upon acquiring $100 in new reserves and deposits? (The answer is not $800!)

return if things go badly minus $90 to repay the loan). That implies that the investor will lose $5 off of the $10 of his own savings that he himself originally put into the investment. That is a 50 percent loss—ten times as much as the 5 percent loss that he would have sustained if he had used only his own money to make the $100 investment. Thus, you can see that leverage increases both profits if things go well and losses if things go badly.

A modern bank uses a lot of leverage. In fact, only about 5 percent of the money that it invests comes from its shareholders and the money they paid to purchase ownership shares in the bank. The other 95 percent comes from borrowing, either by issuing bonds (about 25 percent) or by taking in checking and savings deposits (about 70 percent). It surprises many people, but checking and savings deposits are technically a type of loan made by depositors to their banks. So banks get 70 percent of their leverage and funding from money borrowed from depositors.

The problem with that much leverage is that it takes only very small losses to drive the bank into insolvency and a situation in which it cannot repay all of the money that it has borrowed because the value of the bank's assets has fallen below the value of the bank's liabilities. Consider what would happen if for every $100 invested by the bank, only $94 returned. That by itself is a 6 percent rate of loss. But because the bank is borrowing $95 of every $100 that it invests, it will be driven into insolvency by that six percent loss because it is getting back less than the amount that it borrowed to fund the investment. With the investment returning only $94, there won't be enough money to pay off the $95 that the bank borrowed!

So why do banks use this much leverage? Because it is very profitable for the bankers who run the banks. If things are going well, the leverage massively increases the banks' profit rates—and bankers get paid bonuses based on those profit rates. On the other hand, if things go badly, bankers have come to expect bailouts in which the government uses taxpayer money to ensure that all of a bank's liabilities are repaid. So from the perspective of the bankers, it's "heads I win, tails you lose."

One solution to these problems would be to require banks to use much less leverage. For every $100 that a bank wants to invest, the bank could be required to raise $30 from its shareholder owners so that it would only be borrowing the other $70 through issuing bonds or by taking in checking and savings deposits. That would make the entire banking system much more stable because it would be very unlikely that for any $100 invested into projects by the bank, less than $70 would come back. Even if only $71 came back, there would still be enough money to pay back the $70 of borrowed money—and thus no need for the bank to go bankrupt or require a government bailout via the deposit insurance system.

Unfortunately, however, bankers have lobbied strongly and successfully against any legal attempts to require lower leverage levels. So the current regulatory system relies instead on bank supervisors who attempt to prevent the banks from making bad loans. That system was unable to prevent the 2007–2008 financial crisis and a number of economists argue that until leverage is reduced, no amount of bank supervision will be sufficient to prevent another financial crisis because with massive leverage, even a small loss can destroy a bank.

- Suppose the banking system is loaned up and faces a 20 percent reserve ratio. Explain how it might have to reduce its outstanding loans by $400 when a $100 cash withdrawal from a checkable-deposit account forces one bank to draw down its reserves by $100.

Reversibility: The Multiple Destruction of Money

The process we have described is reversible. Just as checkable-deposit money is created when banks make loans, checkable-deposit money is destroyed when loans are paid off. Loan repayment, in effect, sets off a process of multiple destruction of money the opposite of the multiple creation process. Because loans are both made and paid off in any period, the direction of the loans, checkable deposits, and money supply in a given period will depend on the net effect of the two processes. If the dollar amount of loans made in some period exceeds the dollar amount of loans paid off, checkable deposits will expand and the money supply will increase. But if the dollar amount of loans is less than the dollar amount of loans paid off, checkable deposits will contract and the money supply will decline.

QUICK REVIEW 33.3

- A single bank in a multibank system can safely lend (create money) by an amount equal to its excess reserves; the banking system can lend (create money) by a multiple of its excess reserves.
- The monetary multiplier is the reciprocal of the required reserve ratio; it is the multiple by which the banking system can expand the money supply for each dollar of excess reserves.
- The monetary multiplier works in both directions; it applies to money destruction from the payback of loans as well as the money creation from the making of loans.

SUMMARY

LO33.1 Discuss why the U.S. banking system is called a "fractional reserve" system.

Modern banking systems are fractional reserve systems: Only a fraction of checkable deposits is backed by currency.

LO33.2 Explain the basics of a bank's balance sheet and the distinction between a bank's actual reserves and its required reserves.

The operation of a commercial bank can be understood through its balance sheet, where assets equal liabilities plus net worth.

Commercial banks keep required reserves on deposit in a Federal Reserve Bank or as vault cash. These required reserves are equal to a specified percentage of the commercial bank's checkable-deposit liabilities. Excess reserves are equal to actual reserves minus required reserves.

Banks lose both reserves and checkable deposits when checks are drawn against them.

LO33.3 Describe how a bank can create money.

Commercial banks create money—checkable deposits, or checkable-deposit money—when they make loans. They convert IOUs, which are *not* money, into checkable-deposits, which *are* money. Money is destroyed when lenders repay bank loans.

The ability of a single commercial bank to create money by lending depends on the size of its excess reserves. Generally speaking, a commercial bank can lend only an amount equal to its excess reserves. Money creation is thus limited because, in all likelihood, checks drawn by borrowers will be deposited in other banks, causing a loss of reserves and deposits to the lending bank equal to the amount of money that it has lent.

Rather than making loans, banks may decide to use excess reserves to buy bonds from the public. In doing so, banks merely credit the checkable-deposit accounts of the bond sellers, thus creating checkable-deposit money. Money vanishes when banks sell bonds to the public because bond buyers must draw down their checkable-deposit balances to pay for the bonds.

Banks earn interest by making loans and by purchasing bonds; they maintain liquidity by holding cash and excess reserves. The Fed pays interest on excess reserves. Nevertheless, banks often can obtain higher interest rates by lending out excess reserves on an overnight basis to banks that are short of required reserves. These loans are made in the federal funds market, and the interest paid on the loans is called the federal funds rate.

LO33.4 Describe the multiple expansion of loans and money by the entire banking system.

The commercial banking system as a whole can lend by a multiple of its excess reserves because the system as a whole cannot lose reserves. Individual banks, however, can lose reserves to other banks in the system.

LO33.5 Define the monetary multiplier, explain how to calculate it, and demonstrate its relevance.

The multiple by which the banking system can lend on the basis of each dollar of excess reserves is the reciprocal of the reserve ratio. This multiple credit expansion process is reversible.

TERMS AND CONCEPTS

fractional reserve banking system	required reserves	actual reserves
balance sheet	reserve ratio	federal funds rate
vault cash	excess reserves	monetary multiplier

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Explain why merchants accepted gold receipts as a means of payment even though the receipts were issued by goldsmiths, not the government. What risk did goldsmiths introduce into the payments system by issuing loans in the form of gold receipts? **LO33.1**
2. Why is the banking system in the United States referred to as a fractional reserve bank system? What is the role of deposit insurance in a fractional reserve system? **LO33.1**
3. What is the difference between an asset and a liability on a bank's balance sheet? How does net worth relate to each? Why must a balance sheet always balance? What are the major assets and claims on a commercial bank's balance sheet? **LO33.2**
4. Why does the Federal Reserve require commercial banks to have reserves? Explain why reserves are an asset to commercial banks but a liability to the Federal Reserve Banks. What are excess reserves? How do you calculate the amount of excess reserves held by a bank? What is the significance of excess reserves? **LO33.2**
5. "Whenever currency is deposited in a commercial bank, cash goes out of circulation and, as a result, the supply of money is reduced." Do you agree? Explain why or why not. **LO33.2**
6. "When a commercial bank makes loans, it creates money; when loans are repaid, money is destroyed." Explain. **LO33.3**

7. Suppose that Mountain Star Bank discovers that its reserves will temporarily fall slightly below those legally required. How might it temporarily remedy this situation through the federal funds market? Now assume Mountain Star finds that its reserves will be substantially and permanently deficient. What remedy is available to this bank? (Hint: Recall your answer to discussion question 6.) **LO33.3**
8. Explain why a single commercial bank can safely lend only an amount equal to its excess reserves, but the commercial banking system as a whole can lend by a multiple of its excess reserves. What is the monetary multiplier, and how does it relate to the reserve ratio? **LO33.4, LO33.5**
9. How would a decrease in the reserve requirement affect the (*a*) size of the monetary multiplier, (*b*) amount of excess reserves in the banking system, and (*c*) extent to which the system could expand the money supply through the creation of checkable deposits via loans? **LO33.5**
10. **LAST WORD** Does leverage increase the total size of the gain or loss from an investment, or just the percentage rate of return on the part of the investment amount that was not borrowed? How would lowering leverage make the financial system more stable?

REVIEW QUESTIONS

1. A goldsmith has $2 million of gold in his vaults. He issues $5 million in gold receipts. His gold holdings are what fraction of the paper money (gold receipts) he has issued? **LO33.1**
   a. 1/10.
   b. 1/5.
   c. 2/5.
   d. 5/5.
2. A commercial bank has $100 million in checkable-deposit liabilities and $12 million in actual reserves. The required reserve ratio is 10 percent. How big are the bank's excess reserves? **LO33.2**
   a. $100 million.
   b. $88 million.
   c. $12 million.
   d. $2 million.
3. The actual reason that banks must hold required reserves is: **LO33.2**
   a. To enhance liquidity and deter bank runs.
   b. To help fund the Federal Deposit Insurance Corporation, which insures bank deposits.
   c. To give the Fed control over the lending ability of commercial banks.
   d. To help increase the number of bank loans.
4. A single commercial bank in a multibank banking system can lend only an amount equal to its initial preloan ________________. **LO33.3**
   a. Total reserves.
   b. Excess reserves.
   c. Total deposits.
   d. Excess deposits.
5. The two conflicting goals facing commercial banks are: **LO33.3**
   a. Profit and liquidity.
   b. Profit and loss.
   c. Deposits and withdrawals.
   d. Assets and liabilities.
6. Suppose that the banking system in Canada has a required reserve ratio of 10 percent while the banking system in the United States has a required reserve ratio of 20 percent. In which country would $100 of initial excess reserves be able to cause a larger total amount of money creation? **LO33.4**
   a. Canada.
   b. United States.
7. Suppose that the Fed has set the reserve ratio at 10 percent and that banks collectively have $2 billion in excess reserves. What is the maximum amount of new checkable-deposit money that can be created by the banking system? **LO33.5**
   a. $0.
   b. $200 million.
   c. $2 billion.
   d. $20 billion.
8. Suppose that last year $30 billion in new loans were extended by banks while $50 billion in old loans were paid off by borrowers. What happened to the money supply? **LO33.5**
   a. Increased.
   b. Decreased.
   c. Stayed the same.

PROBLEMS

1. Suppose the assets of the Silver Lode Bank are $100,000 higher than on the previous day and its net worth is up $20,000. By how much and in what direction must its liabilities have changed from the day before? **LO33.2**
2. Suppose that Serendipity Bank has excess reserves of $8,000 and checkable deposits of $150,000. If the reserve ratio is 20 percent, what is the size of the bank's actual reserves? **LO33.2**
3. Third National Bank has reserves of $20,000 and checkable deposits of $100,000. The reserve ratio is 20 percent. Households deposit $5,000 in currency into the bank and that currency is added to reserves. What level of excess reserves does the bank now have? **LO33.3**
4. Suppose again that Third National Bank has reserves of $20,000 and checkable deposits of $100,000. The reserve ratio is 20 percent. The bank now sells $5,000 in securities

Assets		(1)	(2)	Liabilities and net worth		(1′)	(2′)
Reserves	$22,000	_____	_____	Checkable deposits	$100,000	_____	_____
Securities	38,000	_____	_____				
Loans	40,000	_____	_____				

to the Federal Reserve Bank in its district, receiving a $5,000 increase in reserves in return. What level of excess reserves does the bank now have? By what amount does your answer differ (yes, it does!) from the answer to problem 3? **LO33.3**

5. The balance sheet at the top of the page is for Big Bucks Bank. The reserve ratio is 20 percent. **LO33.3**
   a. What is the maximum amount of new loans that Big Bucks Bank can make? Show in columns 1 and 1′ how the bank's balance sheet will appear after the bank has lent this additional amount.
   b. By how much has the supply of money changed?
   c. How will the bank's balance sheet appear after checks drawn for the entire amount of the new loans have been cleared against the bank? Show the new balance sheet in columns 2 and 2′.
   d. Answer questions *a*, *b*, and *c* on the assumption that the reserve ratio is 15 percent.
6. Suppose the simplified consolidated balance sheet shown in the right column is for the entire commercial banking system and that all figures are in billions of dollars. The reserve ratio is 25 percent. **LO33.5**
   a. What is the amount of excess reserves in this commercial banking system? What is the maximum amount the banking system might lend? Show in columns 1 and 1′ how the consolidated balance sheet would look after this amount has been lent. What is the size of the monetary multiplier?

Assets		(1)	Liabilities and net worth		(1′)
Reserves	$ 52	_____	Checkable deposits	$200	_____
Securities	48	_____			
Loans	100	_____			

   b. Answer the questions in part *a* assuming the reserve ratio is 20 percent. What is the resulting difference in the amount that the commercial banking system can lend?
7. If the required reserve ratio is 10 percent, what is the monetary multiplier? If the monetary multiplier is 4, what is the required reserve ratio? **LO33.5**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **34**

Interest Rates and Monetary Policy

Learning Objectives

LO34.1 **Discuss how the equilibrium interest rate is determined in the market for money.**

LO34.2 **Describe the balance sheet of the Federal Reserve and the meaning of its major items.**

LO34.3 **List and explain the goals and tools of monetary policy.**

LO34.4 **Describe the federal funds rate and how the Fed directly influences it.**

LO34.5 **Identify the mechanisms by which monetary policy affects GDP and the price level.**

LO34.6 **Explain the effectiveness of monetary policy and its shortcomings.**

Some newspaper commentators have stated that the chairperson of the Federal Reserve Board (currently Ben Bernanke) is the second most powerful person in the United States, after the U.S. president. That is undoubtedly an exaggeration because the chair has only a single vote on the 7-person Federal Reserve Board and 12-person Federal Open Market Committee. But there can be no doubt about the chair's influence as well as the overall importance of the Federal Reserve and the **monetary policy** that it conducts. Such policy consists of deliberate changes in the money supply to influence interest rates and thus the total level of spending in the economy. The goal of monetary policy is to achieve and maintain price-level stability, full employment, and economic growth.

Interest Rates

LO34.1 Discuss how the equilibrium interest rate is determined in the market for money.

The Fed's primary influence on the economy in normal economic times is through its ability to change the money supply (M_1 and M_2) and therefore affect interest rates. Interest rates can be thought of in several ways. Most basically, **interest** is the price paid for the use of money. It is also the price that borrowers need to pay lenders for transferring purchasing power to the future. And it can be thought of as the amount of money that must be paid for the use of $1 for 1 year. Although there are many different interest rates that vary by purpose, size, risk, maturity, and taxability, we will simply speak of *the* interest rate unless stated otherwise.

Let's see how the interest rate is determined. Because it is a "price," we again turn to demand and supply analysis for the answer.

The Demand for Money

Why does the public want to hold some of its wealth as *money?* There are two main reasons: to make purchases with it and to hold it as an asset.

Transactions Demand, D_t People hold money because it is convenient for purchasing goods and services. Households usually are paid once a week, every 2 weeks, or monthly, whereas their expenditures are less predictable and typically more frequent. So households must have enough money on hand to buy groceries and pay mortgage and utility bills. Nor are business revenues and expenditures simultaneous. Businesses need to have money available to pay for labor, materials, power, and other inputs. The demand for money as a medium of exchange is called the **transactions demand for money.**

The level of nominal GDP is the main determinant of the amount of money demanded for transactions. The larger the total money value of all goods and services exchanged in the economy, the larger the amount of money needed to negotiate those transactions. The transactions demand for money varies directly with nominal GDP. We specify *nominal* GDP because households and firms will want more money for transactions if prices rise or if real output increases. In both instances a larger dollar volume will be needed to accomplish the desired transactions.

In **Figure 34.1a (Key Graph)** we graph the quantity of money demanded for transactions against the interest rate. For simplicity, let's assume that the amount demanded depends exclusively on the level of nominal GDP and is independent of the interest rate. (In reality, higher interest rates are associated with slightly lower volumes of money demanded for transactions.) Our simplifying assumption allows us to graph the transactions demand, D_t, as a vertical line. This demand curve is positioned at $100 billion, on the assumption that each dollar held for transactions purposes is spent an average of three times per year and that nominal GDP is $300 billion. Thus the public needs $100 billion (= $300 billion/3) to purchase that GDP.

Asset Demand, D_a The second reason for holding money derives from money's function as a store of value. People may hold their financial assets in many forms, including corporate stocks, corporate or government bonds, or money. To the extent they want to hold money as an asset, there is an **asset demand for money.**

People like to hold some of their financial assets as money (apart from using it to buy goods and services) because money is the most liquid of all financial assets; it is immediately usable for purchasing other assets when opportunities arise. Money is also an attractive asset to hold when the prices of other assets such as bonds are expected to decline. For example, when the price of a bond falls, the bondholder who sells the bond prior to the payback date of the full principal will suffer a loss (called a *capital loss*). That loss will partially or fully offset the interest received on the bond. Holding money presents no such risk of capital loss from changes in interest rates.

The disadvantage of holding money as an asset is that it earns no or very little interest. Checkable deposits pay either no interest or lower interest rates than bonds. Currency itself earns no interest at all.

Knowing these advantages and disadvantages, the public must decide how much of its financial assets to hold as money, rather than other assets such as bonds. The answer depends primarily on the rate of interest. A household or a business incurs an opportunity cost when it holds money; in both cases, interest income is forgone or sacrificed. If a bond pays 6 percent interest, for example, holding $100 as cash or in a noninterest checkable account costs $6 per year of forgone income.

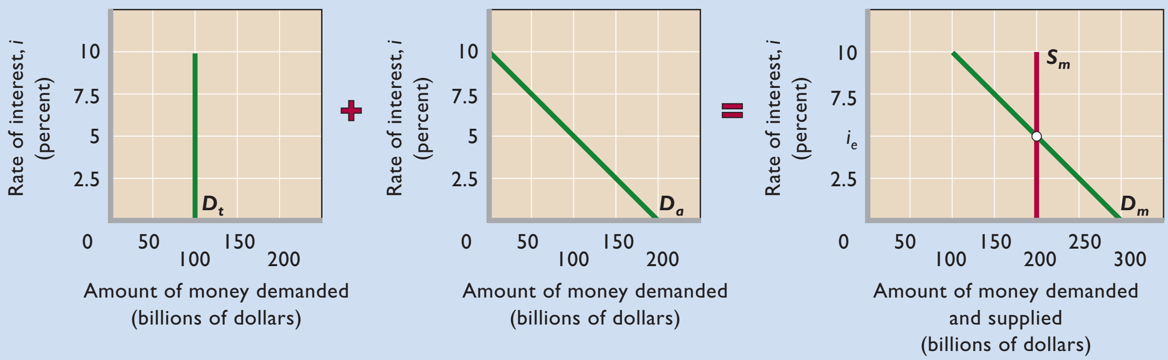
The amount of money demanded as an asset therefore varies inversely with the rate of interest (which is the opportunity cost of holding money as an asset). When the interest rate rises, being liquid and avoiding

ORIGIN OF THE IDEA

O34.1
Liquidity preference

KEY GRAPH

FIGURE 34.1 The demand for money, the supply of money, and the equilibrium interest rate. The total demand for money D_m is determined by horizontally adding the asset demand for money D_a to the transactions demand D_t. The transactions demand is vertical because it is assumed to depend on nominal GDP rather than on the interest rate. The asset demand varies inversely with the interest rate because of the opportunity cost involved in holding currency and checkable deposits that pay no interest or very low interest. Combining the money supply (stock) S_m with the total money demand D_m portrays the market for money and determines the equilibrium interest rate i_e.



(a) Transactions demand for money, D_t

(b) Asset demand for money, D_a

(c) Total demand for money, $D_m = D_t + D_a$, and supply of money, S_m

QUICK QUIZ FOR FIGURE 34.1

1. In this graph, at the interest rate i_e (5 percent):
   - **a.** the amount of money demanded as an asset is \$50 billion.
   - **b.** the amount of money demanded for transactions is \$200 billion.
   - **c.** bond prices will decline.
   - **d.** \$100 billion is demanded for transactions, \$100 billion is demanded as an asset, and the money supply is \$200 billion.
2. In this graph, at an interest rate of 10 percent:
   - **a.** no money will be demanded as an asset.
   - **b.** total money demanded will be \$200 billion.
   - **c.** the Federal Reserve will supply \$100 billion of money.
   - **d.** there will be a \$100 billion shortage of money.
3. Curve D_a slopes downward because:
   - **a.** lower interest rates increase the opportunity cost of holding money.
   - **b.** lower interest rates reduce the opportunity cost of holding money.
   - **c.** the asset demand for money varies directly (positively) with the interest rate.
   - **d.** the transactions-demand-for-money curve is perfectly vertical.
4. Suppose the supply of money declines to \$100 billion. The equilibrium interest rate would:
   - **a.** fall, the amount of money demanded for transactions would rise, and the amount of money demanded as an asset would decline.
   - **b.** rise, and the amounts of money demanded both for transactions and as an asset would fall.
   - **c.** fall, and the amounts of money demanded both for transactions and as an asset would increase.
   - **d.** rise, the amount of money demanded for transactions would be unchanged, and the amount of money demanded as an asset would decline.

Answers: 1. d; 2. a; 3. b; 4. d

capital losses becomes more costly. The public reacts by reducing its holdings of money as an asset. When the interest rate falls, the cost of being liquid and avoiding capital losses also declines. The public therefore increases the amount of financial assets that it wants to hold as money. This inverse relationship just described is shown by D_a in Figure 34.1b.

Total Money Demand, D_m As shown in Figure 34.1, we find the **total demand for money,** D_m, by horizontally adding the asset demand to the transactions demand. The resulting downsloping line in Figure 34.1c represents the total amount of money the public wants to hold, both for transactions and as an asset, at each possible interest rate.

Recall that the transactions demand for money depends on the nominal GDP. A change in the nominal GDP—working through the transactions demand for money—will shift the total money demand curve. Specifically, an increase in nominal GDP means that the public wants to hold a larger amount of money for transactions, and that extra demand will shift the total money demand curve to the right. In contrast, a decline in the nominal GDP will shift the total money demand curve to the left. As an example, suppose nominal GDP increases from \$300 billion to \$450 billion and the average dollar held for transactions is still spent three times per year. Then the transactions demand curve will shift from \$100 billion (= \$300 billion/3) to \$150 billion (= \$450 billion/3). The total money demand curve will then lie \$50 billion farther to the right at each possible interest rate.

WORKED PROBLEMS

W34.1
Demand for money

The Equilibrium Interest Rate

We can combine the demand for money with the supply of money to determine the equilibrium rate of interest. In Figure 34.1c, the vertical line, S_m, represents the money supply. It is a vertical line because the monetary authorities and financial institutions have provided the economy with some particular stock of money. Here it is \$200 billion.

Just as in a product market or a resource market, the intersection of demand and supply determines the equilibrium price in the market for money. In Figure 34.1, this equilibrium price is the equilibrium interest rate, i_e. At this interest rate, the quantity of money demanded (= \$200 billion) equals the quantity of money supplied (= \$200 billion). The equilibrium interest rate can be thought of as the market-determined price that borrowers must pay for using someone else's money over some period of time.

Changes in the demand for money, the supply of money, or both can change the equilibrium interest rate. For reasons that will soon become apparent, we are most interested in changes in the supply of money. The important generalization is this: An increase in the supply of money will lower the equilibrium interest rate; a decrease in the supply of money will raise the equilibrium interest rate.

Interest Rates and Bond Prices

Interest rates and bond prices are inversely related. When the interest rate increases, bond prices fall; when the interest rate falls, bond prices rise. Why so? First understand that bonds are bought and sold in financial markets and that the price of bonds is determined by bond demand and bond supply.

Suppose that a bond with no expiration date pays a fixed \$50 annual interest payment and is selling for its face value of \$1,000. The interest yield on this bond is 5 percent:

$$\frac{\$50}{\$1{,}000} = 5\% \text{ interest yield}$$

Now suppose the interest rate in the economy rises to $7\frac{1}{2}$ percent from 5 percent. Newly issued bonds will pay \$75 per \$1,000 lent. Older bonds paying only \$50 will not be salable at their \$1,000 face value. To compete with the $7\frac{1}{2}$ percent bond, the price of this bond will need to fall to \$667 to remain competitive. The \$50 fixed annual interest payment will then yield $7\frac{1}{2}$ percent to whoever buys the bond:

$$\frac{\$50}{\$667} = 7\tfrac{1}{2}\%$$

Next suppose that the interest rate falls to $2\frac{1}{2}$ percent from the original 5 percent. Newly issued bonds will pay \$25 on \$1,000 loaned. A bond paying \$50 will be highly attractive. Bond buyers will bid up its price to \$2,000, where the yield will equal $2\frac{1}{2}$ percent:

$$\frac{\$50}{\$2{,}000} = 2\tfrac{1}{2}\%$$

The point is that bond prices fall when the interest rate rises and rise when the interest rate falls. There is an inverse relationship between the interest rate and bond prices.

WORKED PROBLEMS

W34.2
Bond prices and interest rates

QUICK REVIEW 34.1

- People demand money for transaction and asset purposes.
- The total demand for money is the sum of the transactions and asset demands; it is graphed as an inverse relationship (downsloping line) between the interest rate and the quantity of money demanded.
- The equilibrium interest rate is determined by money demand and supply; it occurs when people are willing to hold the exact amount of money being supplied by the monetary authorities.
- Interest rates and bond prices are inversely related.

The Consolidated Balance Sheet of the Federal Reserve Banks

LO34.2 Describe the balance sheet of the Federal Reserve and the meaning of its major items.

With this basic understanding of interest rates, we can turn to monetary policy, which relies on changes in interest rates to be effective. The 12 Federal Reserve Banks together constitute the U.S. "central bank," nicknamed the "Fed." (Global Perspective 34.1 also lists some of the other central banks in the world, along with their nicknames.)

The Fed's balance sheet helps us consider how the Fed conducts monetary policy. Table 34.1 consolidates the pertinent assets and liabilities of the 12 Federal Reserve Banks as of April 10, 2013. You will see that some of the Fed's assets and liabilities differ from those found on the balance sheet of a commercial bank.

GLOBAL PERSPECTIVE 34.1

Central Banks, Selected Nations

The monetary policies of the world's major central banks are often in the international news. Here are some of their official names, along with a few of their popular nicknames.

- Australia: Reserve Bank of Australia (RBA)
- Canada: Bank of Canada
- Euro Zone: European Central Bank (ECB)
- Japan: The Bank of Japan (BOJ)
- Mexico: Banco de Mexico (Mex Bank)
- Russia: Central Bank of Russia
- Sweden: Sveriges Riksbank
- United Kingdom: Bank of England
- United States: Federal Reserve System (the "Fed")

Assets

The two main assets of the Federal Reserve Banks are securities and loans to commercial banks. (Again, we will simplify by referring only to *commercial banks*, even though the analysis also applies to *thrifts*—savings and loans, mutual savings banks, and credit unions.)

Securities The securities shown in Table 34.1 are government bonds that have been purchased by the Federal Reserve Banks. They consist largely of Treasury bills (short-term securities), Treasury notes (mid-term securities), and Treasury bonds (long-term securities) issued by the U.S. government to finance past budget deficits. These securities are part of the public debt—the money borrowed by the federal government. The Federal Reserve Banks bought these securities from commercial banks and the public through open-market operations. Although they are an important source of interest income to the Federal Reserve Banks, they are mainly bought and sold to influence the size of commercial bank reserves and, therefore, the ability of those banks to create money by lending.

Loans to Commercial Banks For reasons that will soon become clear, commercial banks occasionally borrow from Federal Reserve Banks. The IOUs that commercial banks give these "bankers' banks" in return for loans are listed on the Federal Reserve balance sheet as "Loans to commercial banks." They are assets to the Fed because they are claims against the commercial banks. To commercial banks, of course, these loans are liabilities in that they must be repaid. Through borrowing in this way, commercial banks can increase their reserves.

Liabilities

On the "liabilities and net worth" side of the Fed's consolidated balance sheet, three entries are noteworthy: reserves, Treasury deposits, and Federal Reserve Notes.

Reserves of Commercial Banks The Fed requires that the commercial banks hold reserves against their

TABLE 34.1 Consolidated Balance Sheet of the 12 Federal Reserve Banks, April 10, 2013 (in Millions)

Assets		Liabilities and Net Worth	
Securities	$2,957,619	Reserves of commercial banks	$1,851,361
Loans to commercial banks	439	Treasury deposits	52,478
All other assets	271,355	Federal Reserve Notes (outstanding)	1,137,087
		All other liabilities and net worth	188,487
Total	$3,229,413	Total	$3,229,413

Source: Federal Reserve Statistical Release, H.4.1, April 10, 2013, **www.federalreserve.gov**.

checkable deposits. The Fed pays interest on these required reserves and also on the excess reserves that banks choose to hold at the Fed. Banks held a huge amount of these excess reserves at the Fed during the severe recession of 2007–2009 and on through at least 2013 as the economy recovered only sluggishly after the Great Recession. Banks simply were highly concerned that loans to some private borrowers might not get paid back. When held in the Federal Reserve Banks, these reserves are listed as a liability on the Fed's balance sheet. They are assets on the books of the commercial banks, which still own them even though they are deposited at the Federal Reserve Banks.

Treasury Deposits The U.S. Treasury keeps deposits in the Federal Reserve Banks and draws checks on them to pay its obligations. To the Treasury these deposits are assets; to the Federal Reserve Banks they are liabilities. The Treasury creates and replenishes these deposits by depositing tax receipts and money borrowed from the public or from the commercial banks through the sale of bonds.

Federal Reserve Notes Outstanding As we have seen, the supply of paper money in the United States consists of Federal Reserve Notes issued by the Federal Reserve Banks. When this money is circulating outside the Federal Reserve Banks, it constitutes claims against the assets of the Federal Reserve Banks. The Fed thus treats these notes as a liability.

QUICK REVIEW 34.2

- The two main assets of the Federal Reserve Banks are securities and loans to commercial banks. Most of the securities are bills, notes, and bonds issued by the U.S. Treasury to finance past federal budget deficits.
- The three major liabilities of the Federal Reserve Banks are reserves of commercial banks, Treasury deposits, and outstanding Federal Reserve notes.

Tools of Monetary Policy

LO34.3 List and explain the goals and tools of monetary policy.

ORIGIN OF THE IDEA

O34.2
Tools of monetary policy

With this look at the Federal Reserve Banks' consolidated balance sheet, we can now explore how the Fed can influence the money-creating abilities of the commercial banking system. The Fed has four main tools of monetary control it can use to alter the reserves of commercial banks:

- Open-market operations
- The reserve ratio
- The discount rate
- Interest on reserves

Open-Market Operations

Bond markets are "open" to all buyers and sellers of corporate and government bonds (securities). The Federal Reserve is the largest single holder of U.S. government securities. The U.S. government, not the Fed, issued these Treasury bills, Treasury notes, and Treasury bonds to finance past budget deficits. Over the decades, the Fed has purchased these securities from major financial institutions that buy and sell government and corporate securities for themselves or their customers.

The Fed's **open-market operations** consist of buying government bonds (U.S. securities) from or selling government bonds to commercial banks and the general public. The conduit for the Fed's open-market operations is the New York Federal Reserve Bank and a group of 21 or so large financial firms called "primary dealers." These financial institutions, in turn, buy the bonds from and sell the bonds to commercial banks and the general public. Open-market operations are the Fed's most important day-to-day instrument for influencing the money supply.

Buying Securities Suppose that the Fed decides to have the Federal Reserve Banks buy government bonds. They can purchase these bonds either from commercial banks or from the public. In both cases the reserves of the commercial banks will increase.

From Commercial Banks When Federal Reserve Banks buy government bonds *from commercial banks*,

(*a*) The commercial banks give up part of their holdings of securities (the government bonds) to the Federal Reserve Banks.

(*b*) The Federal Reserve Banks, in paying for these securities, place newly created reserves in the accounts of the commercial banks at the Fed. (These reserves are created "out of thin air," so to speak!) The reserves of the commercial banks go up by the amount of the purchase of the securities.

We show these outcomes as (*a*) and (*b*) on the following consolidated balance sheets of the commercial banks and the Federal Reserve Banks:

Fed Buys Bonds from Commercial Banks

Federal Reserve Banks

Assets	Liabilities and net worth
+ Securities (*a*)	+ Reserves of commercial banks (*b*)

Commercial Banks

Assets	Liabilities and net worth
− Securities (*a*) + Reserves (*b*)	

The upward arrow shows that securities have moved from the commercial banks to the Federal Reserve Banks. So we enter "− Securities" (minus securities) in the asset column of the balance sheet of the commercial banks. For the same reason, we enter "+ Securities" in the asset column of the balance sheet of the Federal Reserve Banks.

The downward arrow indicates that the Federal Reserve Banks have provided reserves to the commercial banks. So we enter "+ Reserves" in the asset column of the balance sheet for the commercial banks. In the liability column of the balance sheet of the Federal Reserve Banks, the plus sign indicates that although commercial bank reserves have increased, they are a liability to the Federal Reserve Banks because the reserves are owned by the commercial banks.

What is most important about this transaction is that when Federal Reserve Banks purchase securities from commercial banks, they increase the reserves in the banking system, which then increases the lending ability of the commercial banks.

From the Public The effect on commercial bank reserves is much the same when Federal Reserve Banks purchase securities from the general public. Suppose the Gristly Meat Packing Company has government bonds that it sells in the open market to the Federal Reserve Banks. The transaction has several elements:

(*a*) Gristly gives up securities to the Federal Reserve Banks and gets in payment a check drawn by the Federal Reserve Banks on themselves.

(*b*) Gristly promptly deposits the check in its account with the Wahoo bank.

(*c*) The Wahoo bank sends this check against the Federal Reserve Banks to a Federal Reserve Bank for collection. As a result, the Wahoo bank enjoys an increase in its reserves.

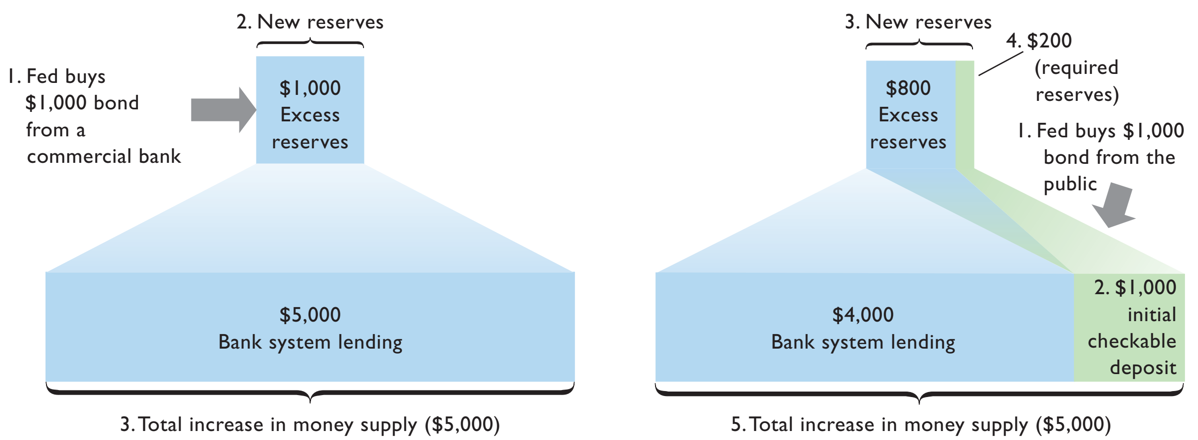
To keep things simple, we will dispense with showing the balance sheet changes resulting from the Fed's sale or purchase of bonds from the public. But two aspects of this transaction are particularly important. First, as with Federal Reserve purchases of securities directly from commercial banks, the purchases of securities from the public increase the lending ability of the commercial banking system. Second, the supply of money is directly increased by the Federal Reserve Banks' purchase of government bonds (aside from any expansion of the money supply that may occur from the increase in commercial bank reserves). This direct increase in the money supply has taken the form of an increased amount of checkable deposits in the economy as a result of Gristly's deposit.

The Federal Reserve Banks' purchases of securities from the commercial banking system differ slightly from their purchases of securities from the public. If we assume that all commercial banks are loaned up initially, Federal Reserve bond purchases *from commercial banks* increase the actual reserves and excess reserves of commercial banks by the entire amount of the bond purchases. As shown in the left panel in Figure 34.2, a $1,000 bond purchase from a commercial bank increases both the actual and the excess reserves of the commercial bank by $1,000.

In contrast, Federal Reserve Bank purchases of bonds from the public increase actual reserves but also increase checkable deposits when the sellers place the Fed's check into their personal checking accounts. Thus, a $1,000 bond purchase from the public would increase checkable deposits by $1,000 and hence the actual reserves of the loaned-up banking system by the same amount. But with a 20 percent reserve ratio applied to the $1,000 checkable deposit, the excess reserves of the banking system would be only $800 since $200 of the $1,000 would have to be held as reserves.

However, in both transactions the end result is the same: When Federal Reserve Banks buy securities in the open market, commercial banks' reserves are increased. When the banks lend out an amount equal to their excess reserves, the nation's money supply will rise. Observe in Figure 34.2 that a $1,000 purchase of

FIGURE 34.2 The Federal Reserve's purchase of bonds and the expansion of the money supply. Assuming all banks are loaned up initially, a Federal Reserve purchase of a $1,000 bond from either a commercial bank or the public can increase the money supply by $5,000 when the reserve ratio is 20 percent. In the left panel of the diagram, the purchase of a $1,000 bond from a commercial bank creates $1,000 of excess reserves that support a $5,000 expansion of checkable deposits through loans. In the right panel, the purchase of a $1,000 bond from the public creates a $1,000 checkable deposit but only $800 of excess reserves because $200 of reserves is required to "back up" the $1,000 new checkable deposit. The commercial banks can therefore expand the money supply by only $4,000 by making loans. This $4,000 of checkable-deposit money plus the new checkable deposit of $1,000 equals $5,000 of new money.



WORKED PROBLEMS

W34.3
Open-market operations

bonds by the Federal Reserve results in a potential of $5,000 of additional money, regardless of whether the purchase was made from commercial banks or from the general public.

Selling Securities As you may suspect, when the Federal Reserve Banks sell government bonds, commercial banks' reserves are reduced. Let's see why.

To Commercial Banks When the Federal Reserve Banks sell securities in the open market to commercial banks,

(*a*) The Federal Reserve Banks give up securities that the commercial banks acquire.

(*b*) The commercial banks pay for those securities by drawing checks against their deposits—that is, against their reserves—in Federal Reserve Banks. The Fed collects on those checks by reducing the commercial banks' reserves accordingly.

The balance-sheet changes—again identified by (*a*) and (*b*)—appear as shown in the following balance sheets. The reduction in commercial bank reserves is indicated by the minus signs before the appropriate entries.

Fed Sells Bonds to Commercial Banks

Federal Reserve Banks

Assets	Liabilities and net worth
− Securities (*a*)	− Reserves of commercial banks (*b*)
(*a*) Securities	(*b*) Reserves

Commercial Banks

Assets	Liabilities and net worth
− Reserves (*b*) + Securities (*a*)	

To the Public When the Federal Reserve Banks sell securities to the public, the outcome is much the same. Let's put the Gristly Company on the buying end of government bonds that the Federal Reserve Banks are selling:

(*a*) The Federal Reserve Banks sell government bonds to Gristly, which pays with a check drawn on the Wahoo bank.

(*b*) The Federal Reserve Banks clear this check against the Wahoo bank by reducing Wahoo's reserves.

TABLE 34.2 The Effects of Changes in the Reserve Ratio on the Lending Ability of Commercial Banks

(1) Reserve Ratio, %	(2) Checkable Deposits	(3) Actual Reserves	(4) Required Reserves	(5) Excess Reserves, (3) − (4)	(6) Money-Creating Potential of Single Bank, = (5)	(7) Money-Creating Potential of Banking System
(1) 10	$20,000	$5,000	$2,000	$ 3,000	$ 3,000	$ 30,000
(2) 20	20,000	5,000	4,000	1,000	1,000	5,000
(3) 25	20,000	5,000	5,000	0	0	0
(4) 30	20,000	5,000	6,000	−1,000	−1,000	−3,333

(*c*) The Wahoo bank returns the canceled check to Gristly, reducing Gristly's checkable deposit accordingly.

Federal Reserve bond sales of $1,000 to the commercial banking system reduce the system's actual and excess reserves by $1,000. But a $1,000 bond sale to the public reduces excess reserves by $800 because the public's checkable-deposit money is also reduced by $1,000 by the sale. Since the commercial banking system's outstanding checkable deposits are reduced by $1,000, banks need keep $200 less in reserves.

Whether the Fed sells bonds to the public or to commercial banks, the result is the same: When Federal Reserve Banks sell securities in the open market, commercial bank reserves are reduced. If all excess reserves are already lent out, this decline in commercial bank reserves produces a decline in the nation's money supply. In our example, a $1,000 sale of government securities results in a $5,000 decline in the money supply whether the sale is made to commercial banks or to the general public. You can verify this by reexamining Figure 34.2 and tracing the effects of a *sale* of a $1,000 bond by the Fed either to commercial banks or to the public.

What makes commercial banks and the public willing to sell government securities to, or buy them from, Federal Reserve Banks? The answer lies in the price of bonds and their interest yields. We know that bond prices and interest rates are inversely related. When the Fed buys government bonds, the demand for them increases. Government bond prices rise, and their interest yields decline. The higher bond prices and their lower interest yields prompt banks, securities firms, and individual holders of government bonds to sell them to the Federal Reserve Banks.

When the Fed sells government bonds, the additional supply of bonds in the bond market lowers bond prices and raises their interest yields, making government bonds attractive purchases for banks and the public.

The Reserve Ratio

The Fed also can manipulate the **reserve ratio** in order to influence the ability of commercial banks to lend. Suppose a commercial bank's balance sheet shows that reserves are $5,000 and checkable deposits are $20,000. If the legal reserve ratio is 20 percent (row 2, Table 34.2), the bank's required reserves are $4,000. Since actual reserves are $5,000, the excess reserves of this bank are $1,000. On the basis of $1,000 of excess reserves, this one bank can lend $1,000; however, the banking system as a whole can create a maximum of $5,000 of new checkable-deposit money by lending (column 7).

Raising the Reserve Ratio Now, what if the Fed raised the reserve ratio from 20 to 25 percent? (See row 3.) Required reserves would jump from $4,000 to $5,000, shrinking excess reserves from $1,000 to zero. Raising the reserve ratio increases the amount of required reserves banks must keep. As a consequence, either banks lose excess reserves, diminishing their ability to create money by lending, or they find their reserves deficient and are forced to contract checkable deposits and therefore the money supply. In the example in Table 34.2, excess reserves are transformed into required reserves, and the money-creating potential of our single bank is reduced from $1,000 to zero (column 6). Moreover, the banking system's money-creating capacity declines from $5,000 to zero (column 7).

What if the Fed increases the reserve requirement to 30 percent? (See row 4.) The commercial bank, to protect itself against the prospect of failing to meet this requirement, would be forced to lower its checkable deposits and at the same time increase its reserves. To reduce its checkable deposits, the bank could let outstanding loans mature and be repaid without extending new credit. To increase reserves, the bank might sell some of its bonds, adding the proceeds to its reserves. Both actions would reduce the supply of money.

Lowering the Reserve Ratio What would happen if the Fed lowered the reserve ratio from the original 20 percent to 10 percent? (See row 1.) In this case, required reserves would decline from $4,000 to $2,000, and excess reserves would jump from $1,000 to $3,000. The single bank's lending (money-creating) ability would increase from $1,000 to $3,000 (column 6), and the banking system's money-creating potential would expand from $5,000 to $30,000 (column 7). Lowering the reserve ratio transforms required reserves into excess reserves and enhances the ability of banks to create new money by lending.

The examples in Table 34.2 show that a change in the reserve ratio affects the money-creating ability of the *banking system* in two ways:

- It changes the amount of excess reserves.
- It changes the size of the monetary multiplier.

For example, when the legal reserve ratio is raised from 10 to 20 percent, excess reserves are reduced from $3,000 to $1,000 and the checkable-deposit multiplier is reduced from 10 to 5. The money-creating potential of the banking system declines from $30,000 (= $3,000 × 10) to $5,000 (= $1,000 × 5). Raising the reserve ratio forces banks to reduce the amount of checkable deposits they create through lending.

The Discount Rate

One of the functions of a central bank is to be a "lender of last resort." Occasionally, commercial banks have unexpected and immediate needs for additional funds. In such cases, each Federal Reserve Bank will make short-term loans to commercial banks in its district.

When a commercial bank borrows, it gives the Federal Reserve Bank a promissory note (IOU) drawn against itself and secured by acceptable collateral—typically U.S. government securities. Just as commercial banks charge interest on the loans they make to their clients, so too Federal Reserve Banks charge interest on loans they grant to commercial banks. The interest rate they charge is called the **discount rate.**

As a claim against the commercial bank, the borrowing bank's promissory note is an asset to the lending Federal Reserve Bank and appears on its balance sheet as "Loans to commercial banks." To the commercial bank the IOU is a liability, appearing as "Loans from the Federal Reserve Banks" on the commercial bank's balance sheet. [See the two (*a*) entries on the balance sheets that follow.]

Commercial Bank Borrowing from the Fed

Federal Reserve Banks

Assets	Liabilities and net worth
+ Loans to commercial banks (*a*)	+ Reserves of commercial banks (*b*)
IOUs	+ Reserves

Commercial Banks

Assets	Liabilities and net worth
+ Reserves (*b*)	+ Loans from the Federal Reserve Banks (*a*)

In providing the loan, the Federal Reserve Bank increases the reserves of the borrowing commercial bank. Since no required reserves need be kept against loans from Federal Reserve Banks, all new reserves acquired by borrowing from Federal Reserve Banks are excess reserves. [These changes are reflected in the two (*b*) entries on the balance sheets.]

In short, borrowing from the Federal Reserve Banks by commercial banks increases the reserves of the commercial banks and enhances their ability to extend credit.

The Fed has the power to set the discount rate at which commercial banks borrow from Federal Reserve Banks. From the commercial banks' point of view, the discount rate is a cost of acquiring reserves. A lowering of the discount rate encourages commercial banks to obtain additional reserves by borrowing from Federal Reserve Banks. When the commercial banks lend new reserves, the money supply increases.

An increase in the discount rate discourages commercial banks from obtaining additional reserves through borrowing from the Federal Reserve Banks. So the Fed may raise the discount rate when it wants to restrict the money supply.

Interest on Reserves

In 2008, federal law was changed so that the Federal Reserve could for the first time pay banks **interest on reserves** held at the Fed. Before that time, any reserves held on deposit at the Federal Reserve were paid zero interest. Thus, before 2008, banks had an incentive to keep their reserves as small as possible because any money kept on reserve at the Fed was money earning a zero percent rate of return for the banks.

The Fed's newfound ability to pay interest on reserves provided the Fed with a fourth policy tool by which it can implement monetary policy and either increase or decrease the amount of monetary stimulus in the economy. As an example, suppose that the Fed wishes to *reduce* the amount of bank lending and, consequently, the amount of money circulating in the economy. It can do so by increasing the rate of interest that it pays on reserves held at the Fed. The higher that interest rate, the more of an incentive banks will have to reduce their risky commercial lending for car, mortgage, and business loans in order to instead increase the reserves that they hold at the Fed and thereby earn the risk-free interest rate that the Fed is paying on reserves.

By contrast, if the Fed wishes to *increase* the amount of money that banks lend into the economy, the Fed can lower the interest rate that it pays on reserves. The lower rate will make it less attractive for banks to keep reserves, and, consequently, banks will be incentivized to increase consumer and commercial lending and thereby stimulate the economy.

In 2012, the rate of interest on reserves was 0.25 percent per year. Given that banks held about $1.7 trillion in reserves at the Fed that year, the Fed paid a total of about $4.25 billion (= 0.0025 times $1.7 trillion) in interest payments to banks for reserves held at the Fed in 2012.

Relative Importance

All four of the Fed's instruments of monetary control are useful in particular economic circumstances, but open-market operations are clearly the most important of the four tools over the course of the business cycle. The buying and selling of securities in the open market has the advantage of flexibility—government securities can be purchased or sold daily in large or small amounts—and the impact on bank reserves is prompt. And, compared with reserve-requirement changes, open-market operations work subtly and less directly. Furthermore, the ability of the Federal Reserve Banks to affect commercial bank reserves through the purchase and sale of bonds is virtually unquestionable. The Federal Reserve Banks have very large holdings of government securities ($2,958 billion in early 2013, for example). The sale of those securities could theoretically reduce commercial bank reserves to zero.

Changing the reserve requirement is a potentially powerful instrument of monetary control, but the Fed has used this technique only sparingly. Normally, it can accomplish its monetary goals more easily through open-market operations. The last change in the reserve requirement was in 1992, when the Fed reduced the requirement from 12 percent to 10 percent. The main purpose was to shore up the profitability of banks and thrifts in the aftermath of the 1990–1991 recession rather than to reduce interest rates by increasing reserves and expanding the money supply.

Until recently, the discount rate was mainly a passive tool of monetary control, with the Fed raising and lowering the rate simply to keep it in line with other interest rates. However, during the financial crisis of 2007–2008, the Fed aggressively lowered the discount rate independently of other interest rates to provide a cheap and plentiful source of reserves to banks whose reserves were being sharply reduced by unexpectedly high default rates on home mortgage loans. Banks borrowed billions at the lower discount rate. This allowed them to meet reserve ratio requirements and thereby preserved their ability to keep extending loans.

QUICK REVIEW 34.3

- The Fed has four main tools of monetary control, each of which works by changing the amount of reserves in the banking system: (a) conducting open-market operations (the Fed's buying and selling of government bonds to the banks and the public); (b) changing the reserve ratio (the percentage of commercial bank deposit liabilities required as reserves); (c) changing the discount rate (the interest rate the Federal Reserve Banks charge on loans to banks and thrifts); and (d) changing the interest rate that it pays on reserves held at the Fed.
- Open-market operations are the Fed's monetary control mechanism of choice for routine increases or decreases in bank reserves over the business cycle; in contrast, changes in reserve requirements and aggressive changes in discount rates or interest on reserves are used only in special situations.

Targeting the Federal Funds Rate

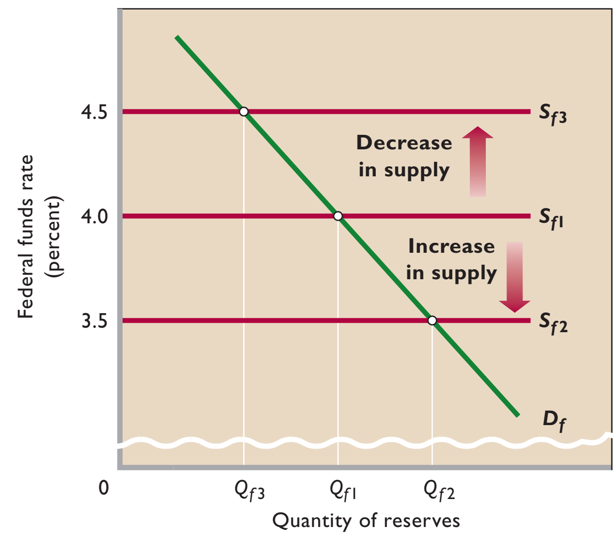
LO34.4 Describe the federal funds rate and how the Fed directly influences it.

The Federal Reserve focuses monetary policy on the interest rate that it can directly influence: the **federal funds rate.** From the previous chapter, you know that this is the rate of interest that banks charge one another on overnight loans made from temporary excess reserves. Recall that the Federal Reserve requires banks (and thrifts) to deposit in their regional Federal Reserve Bank a certain percentage of their checkable deposits as reserves. At the end of any business day, some banks temporarily have excess reserves (more actual reserves than required) and other banks have reserve deficiencies (fewer reserves than required). Because reserves held at the Federal Reserve Banks earn less interest than commercial banks can obtain from overnight loans to other banks, banks with excess reserves usually desire to make such loans to other banks that temporarily need them to meet their reserve requirements. The funds being lent and borrowed overnight are called "federal funds" because they are reserves (funds) that are required by the Federal Reserve to meet reserve requirements. An equilibrium interest rate—the federal funds rate—arises in this market for bank reserves.

The Federal Reserve targets the federal funds rate by manipulating the supply of reserves that are offered in the federal funds market. As previously explained, by buying and selling government bonds, the Fed can increase or decrease the reserves in the banking system. These changes in total reserves in turn affect the amount of *excess reserves* that are available for supply to the federal funds market by whichever banks end up with them on a given day. For instance, suppose that the level of loans and checkable deposits at Wahoo bank are constant on a certain day. If the Fed then engages in open-market operations such that Wahoo's total reserves increase, Wahoo will find that it has excess reserves. It will want to loan out these excess reserves to bank customers as soon as possible. But in the meanwhile it will supply these funds overnight in the federal funds market.

The Federal Open Market Committee (FOMC) meets regularly to choose a desired federal funds rate. It then directs the Federal Reserve Bank of New York to undertake whatever open-market operations may be necessary to achieve and maintain the targeted rate. We demonstrate how this works in Figure 34.3, where we initially assume the Fed desires a 4 percent interest rate. The demand curve for federal funds, D_f, is downsloping because lower interest rates give the banks with reserve deficiencies a greater incentive to borrow federal funds rather than reduce loans as a way to meet their reserve requirements. The supply curve for federal funds, S_{f1}, is somewhat unusual. Specifically, it is horizontal at the targeted federal funds rate, here 4 percent. (Disregard supply curves S_{f2} and S_{f3} for now.) It is horizontal because the Fed uses open-market operations to manipulate the supply of federal funds so that the quantity supplied of federal funds will exactly equal the quantity demanded of federal funds at the targeted interest rate.

FIGURE 34.3 Targeting the federal funds rate. In implementing monetary policy, the Federal Reserve determines a desired federal funds rate and then uses open-market operations (buying and selling of U.S. securities) to add or subtract bank reserves to achieve and maintain that targeted rate. In an expansionary monetary policy, the Fed increases the supply of reserves, for example, from S_{f1} to S_{f2} in this case, to move the federal funds rate from 4 percent to 3.5 percent. In a restrictive monetary policy, it decreases the supply of reserves, say, from S_{f1} to S_{f3}. Here, the federal funds rate rises from 4 percent to 4.5 percent.

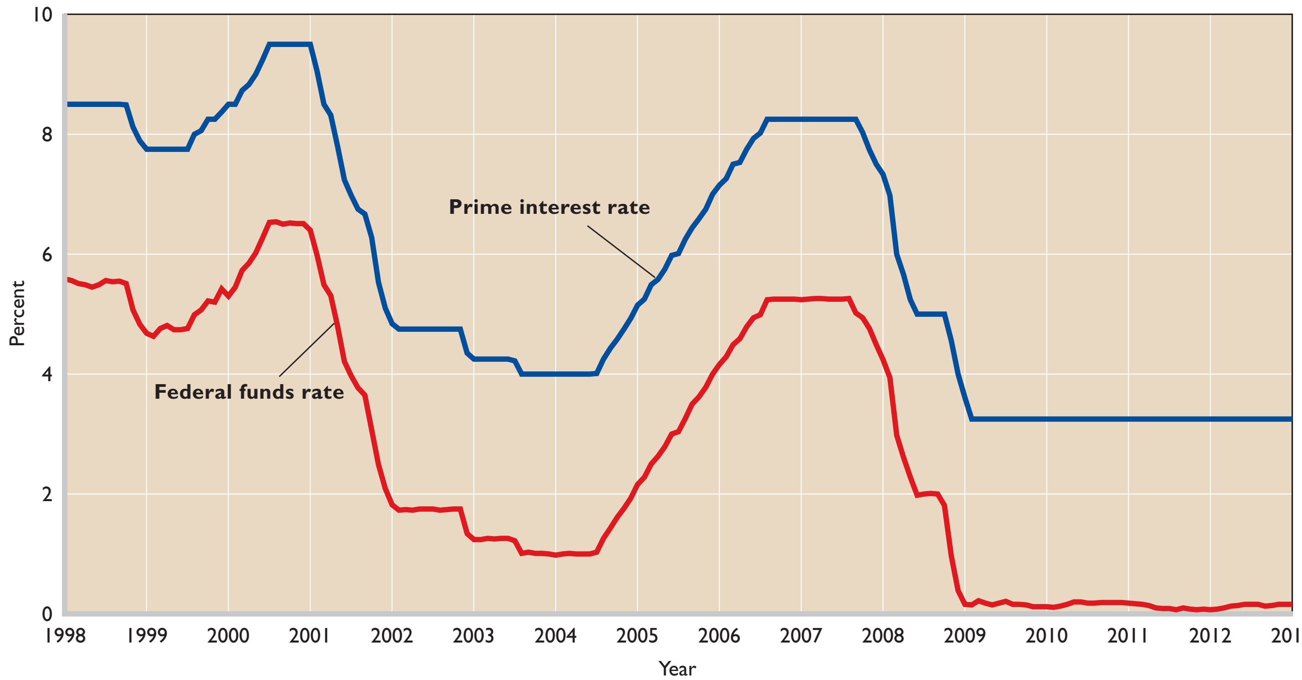


In this case, the Fed seeks to achieve an equilibrium federal funds rate of 4 percent. In Figure 34.3 it is successful. Note that at the 4 percent federal funds rate, the quantity of federal funds supplied (Q_{f1}) equals the quantity of funds demanded (also Q_{f1}). This 4 percent federal funds rate will remain, as long as the supply curve of federal funds is horizontal at 4 percent. If the demand for federal funds increases (D_f shifts to the right along S_{f1}), the Fed will use its open-market operations to increase the availability of reserves such that the 4 percent federal funds rate is retained. If the demand for federal funds declines (D_f shifts to the left along S_{f1}), the Fed will withdraw reserves to keep the federal funds rate at 4 percent. In the language of Chapter 6, the Fed ensures that the supply curve is perfectly elastic at its targeted rate.

Expansionary Monetary Policy

Suppose that the economy faces recession and unemployment. How will the Fed respond? It will initiate an **expansionary monetary policy** (or "easy money policy"). This policy will lower the interest rate to bolster borrowing and spending, which will increase aggregate demand and expand real output. The Fed's immediate step will be to announce a lower target for the federal funds rate, say 3.5 percent instead of 4 percent. To achieve that lower

FIGURE 34.4 The prime interest rate and the federal funds rate in the United States, 1998–2013. The prime interest rate rises and falls with changes in the federal funds rate.



Source: Federal Reserve Statistical Release, Historical Data, H.15, **www.federalreserve.gov.**

rate, the Fed will use open-market operations to buy bonds from banks and the public. We know from previous discussion that the purchase of bonds increases the reserves in the banking system. Alternatively, the Fed could expand reserves by lowering the reserve requirement or lowering the discount rate, but these alternative tools are less frequently used than open-market operations. The Fed could also lower the interest rate that it pays on reserves so that banks will have more of an incentive to lend out reserves rather than keeping them on deposit at the Fed. But through 2013, the Fed had never changed the interest rate that it had paid on reserves after starting to pay interest on reserves in 2008. It remained at 0.25 percent.

The greater reserves in the banking system produce two critical results:

- The supply of federal funds increases, lowering the federal funds rate to the new targeted rate. We show this in Figure 34.3 as a downward shift to the horizontal supply curve from S_{f1} to S_{f2}. The equilibrium federal funds rate falls to 3.5 percent, just as the FOMC wanted. The equilibrium quantity of reserves in the overnight market for reserves rises from Q_{f1} to Q_{f2}.
- A multiple expansion of the nation's money supply occurs (as we demonstrated in Chapter 33). Given the demand for money, the larger money supply places a downward pressure on other interest rates.

One such rate is the **prime interest rate**—the benchmark interest rate used by banks as a reference point for a wide range of interest rates charged on loans to businesses and individuals. The prime interest rate is higher than the federal funds rate because the prime rate involves longer, more risky loans than overnight loans between banks. But the federal funds rate and the prime interest rate closely track one another, as evident in Figure 34.4. Also evident are the changes in these rates over the period shown. We will address these changes later in our discussion of recent monetary policy.

Restrictive Monetary Policy

The opposite monetary policy is in order for periods of rising inflation. The Fed will then undertake a **restrictive monetary policy** (or "tight money policy"). This policy will increase the interest rate to reduce borrowing and spending, which will curtail the expansion of aggregate demand and hold down price-level increases. The Fed's immediate step will be to announce a higher target for the federal funds rate, say 4.5 percent instead of 4 percent.

Through open-market operations, the Fed will sell bonds to the banks and the public and the sale of those bonds will absorb reserves in the banking system. Alternatively, the Fed could absorb reserves by raising the reserve requirement, raising the discount rate, or raising the interest rate that it pays on reserves. But open-market operations are usually sufficient to accomplish the goal.

The smaller reserves in the banking system produce two results opposite those discussed for an expansionary monetary policy:

- The supply of federal funds decreases, raising the federal funds rate to the new targeted rate. We show this in Figure 34.3 as an upward shift of the horizontal supply curve from S_{f1} to S_{f3}. The equilibrium federal funds rate rises to 4.5 percent, just as the FOMC wanted, and the equilibrium quantity of funds in this market falls to Q_{f3}.
- A multiple contraction of the nation's money supply occurs (as demonstrated in Chapter 33). Given the demand for money, the smaller money supply places an upward pressure on other interest rates. For example, the prime interest rate rises.

CONSIDER THIS . . .

The Fed as a Sponge

A good way to remember the role of the Fed in setting the federal funds rate might be to imagine a bowl of water, with the amount of water in the bowl representing the stock of reserves in the banking system. Then think of the FOMC as having a large sponge, labeled open-market operations. When it wants to decrease the federal funds rate, it uses the sponge—soaked with water (reserves) created by the Fed—to squeeze new reserves into the banking system bowl. It continues this process until the higher supply of reserves reduces the federal funds rate to the Fed's desired level. If the Fed wants to increase the federal funds rate, it uses the sponge to absorb reserves from the bowl (banking system). As the supply of reserves falls, the federal funds rate rises to the Fed's desired level.

The Taylor Rule

The proper federal funds rate for a certain period is a matter of policy discretion by the members of the FOMC. At each of their meetings, committee members assess whether the current target for the federal funds rate remains appropriate for achieving the twin goals of low inflation and full employment. If the majority of the FOMC members conclude that a change in the rate is needed, the FOMC sets a new targeted rate. This new target is established without adhering to any particular "inflationary target" or "monetary policy rule." Instead, the committee targets the federal funds rate at the level most appropriate for the current underlying economic conditions.

A rule of thumb suggested by economist John Taylor, however, roughly matches the actual policy of the Fed over many time periods. This rule of thumb builds on the belief held by many economists that central banks are willing to tolerate a small positive rate of inflation if doing so will help the economy to produce at potential output. The **Taylor rule** assumes that the Fed has a 2 percent "target rate of inflation" that it is willing to tolerate and that the FOMC follows three rules when setting its target for the federal funds rate:

- When real GDP equals potential GDP and inflation is at its target rate of 2 percent, the federal funds target rate should be 4 percent, implying a real federal funds rate of 2 percent (= 4 percent nominal federal funds rate *minus* 2 percent inflation rate).
- For each 1 percent increase of real GDP above potential GDP, the Fed should raise the *real* federal funds rate by $\frac{1}{2}$ percentage point.
- For each 1 percent increase in the inflation rate above its 2 percent target rate, the Fed should raise the *real* federal funds rate by $\frac{1}{2}$ percentage point. (Note, though, that in this case each $\frac{1}{2}$ percentage point increase in the real rate will require a 1.5 percentage point increase in the nominal rate to account for the underlying 1 percent increase in the inflation rate.)

The last two rules are applied independently of each other so that if real GDP is above potential output and at the same time inflation is above the 2 percent target rate, the Fed will apply both rules and raise real interest rates in response to both factors. For instance, if real GDP is 1 percent above potential output and inflation is simultaneously 1 percent above the 2 percent target rate, then the Fed will raise the *real* federal funds rate by 1 percentage point (= $\frac{1}{2}$ percentage point for

WORKED PROBLEMS

W34.4

Taylor rule

the excessive GDP + $\frac{1}{2}$ percentage point for the excessive inflation).

Also notice that the last two rules are reversed for situations in which real GDP falls below potential GDP or inflation falls below 2 percent. Each 1 percent decline in real GDP below potential GDP or fall in inflation below 2 percent calls for a decline of the *real* federal funds rate by $\frac{1}{2}$ percentage point.

ORIGIN OF THE IDEA

O34.3

Taylor rule

We reemphasize that the Fed has no official allegiance to the Taylor rule. It changes the federal funds rate to any level that it deems appropriate. During some periods, its policy has diverged significantly from the Taylor rule.

QUICK REVIEW 34.4

- The Fed conducts its monetary policy by establishing a targeted federal funds interest rate—the rate that commercial banks charge one another for overnight loans of reserves.
- An expansionary monetary policy (loose money policy) lowers the federal funds rate, increases the money supply, and lowers other interest rates.
- A restrictive monetary policy (tight money policy) increases the federal funds rate, reduces the money supply, and increases other interest rates.
- The Fed uses it discretion in setting the federal funds target rate, but its decisions regarding monetary policy and the target rate appear to be broadly consistent with the Taylor rule over many time periods.

Monetary Policy, Real GDP, and the Price Level

LO34.5 Identify the mechanisms by which monetary policy affects GDP and the price level.

We have identified and explained the tools of expansionary and contractionary monetary policy. We now want to emphasize how monetary policy affects the economy's levels of investment, aggregate demand, real GDP, and prices.

Cause-Effect Chain

The four diagrams in **Figure 34.5 (Key Graph)** will help you understand how monetary policy works toward achieving its goals.

Market for Money Figure 34.5a represents the market for money, in which the demand curve for money and the supply curve of money are brought together. Recall that the total demand for money is made up of the transactions and asset demands.

This figure also shows three potential money supply curves, S_{m1}, S_{m2}, and S_{m3}. In each case, the money supply is shown as a vertical line representing some fixed amount of money determined by the Fed.

The equilibrium interest rate is the rate at which the amount of money demanded and the amount supplied are equal. With money demand D_m in Figure 34.5a, if the supply of money is \$125 billion (S_{m1}), the equilibrium interest rate is 10 percent. With a money supply of \$150 billion (S_{m2}), the equilibrium interest rate is 8 percent; with a money supply of \$175 billion (S_{m3}), it is 6 percent.

You know from Chapter 28 that the real, not the nominal, rate of interest is critical for investment decisions. So here we assume that Figure 34.5a portrays real interest rates.

Investment These 10, 8, and 6 percent real interest rates are carried rightward to the investment demand curve in Figure 34.5b. This curve shows the inverse relationship between the interest rate—the cost of borrowing to invest—and the amount of investment spending. At the 10 percent interest rate, it will be profitable for the nation's businesses to invest \$15 billion; at 8 percent, \$20 billion; at 6 percent, \$25 billion.

Changes in the interest rate mainly affect the investment component of total spending, although they also affect spending on durable consumer goods (such as autos) that are purchased on credit. The impact of changing interest rates on investment spending is great because of the large cost and long-term nature of capital purchases. Capital equipment, factory buildings, and warehouses are tremendously expensive. In absolute terms, interest charges on funds borrowed for these purchases are considerable.

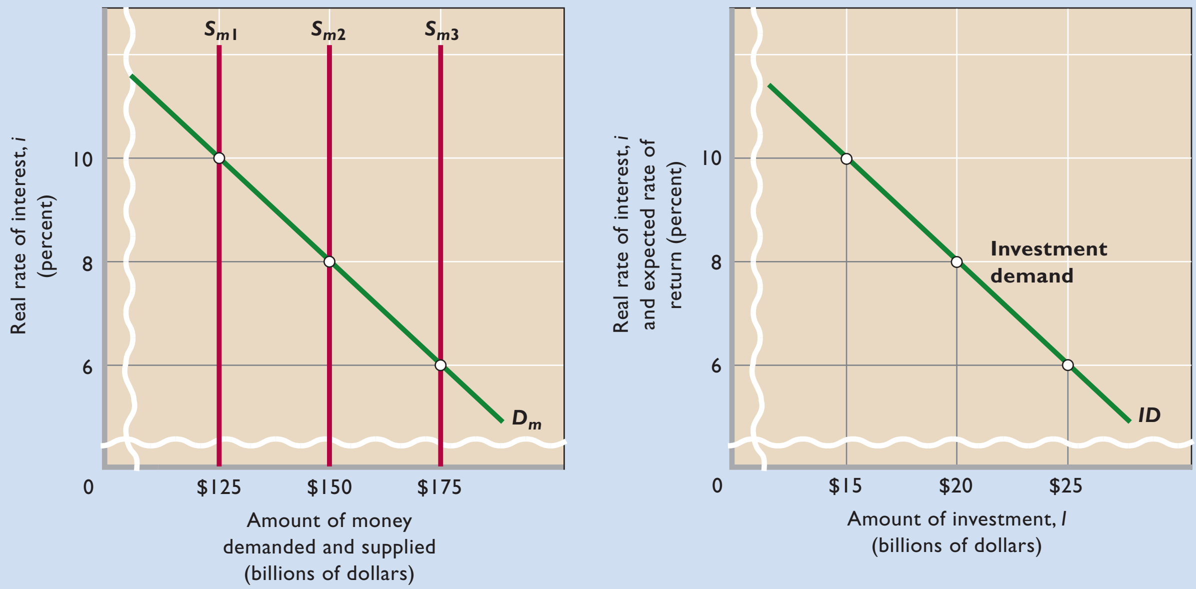
Similarly, the interest cost on a house purchased on a long-term contract is very large: A $\frac{1}{2}$-percentage-point change in the interest rate could amount to thousands of dollars in the total cost of buying a home.

In brief, the impact of changing interest rates is mainly on investment (and, through that, on aggregate demand, output, employment, and the price level). Moreover, as Figure 34.5b shows, investment spending varies inversely with the real interest rate.

Equilibrium GDP Figure 34.5c shows the impact of our three real interest rates and corresponding levels of

KEY GRAPH

FIGURE 34.5 Monetary policy and equilibrium GDP. An expansionary monetary policy that shifts the money supply curve rightward from S_{m1} to S_{m2} in (a) lowers the interest rate from 10 to 8 percent in (b). As a result, investment spending increases from \$15 billion to \$20 billion, shifting the aggregate demand curve rightward from AD_1 to AD_2 in (c) so that real output rises from the recessionary level of \$880 billion to the full employment level Q_f = \$900 billion along the horizontal dashed line. In (d), the economy at point *a* has an inflationary output gap of \$10 billion because it is producing at \$910 billion, \$10 billion above potential output. A restrictive monetary policy that shifts the money supply curve leftward from S_{m3} = \$175 billion to just \$162.5 billion in (a) will increase the interest rate from 6 percent to 7 percent. Investment spending thus falls by \$2.5 billion from \$25 billion to \$22.5 billion in (b). This initial decline is multiplied by 4 by the multiplier process so that the aggregate demand curve shifts leftward in (d) by \$10 billion from AD_3 to AD_4, moving the economy along the horizontal dashed line to equilibrium *b*. This returns the economy to full employment output and eliminates the inflationary output gap.



(a)
The market for money

(b)
Investment demand

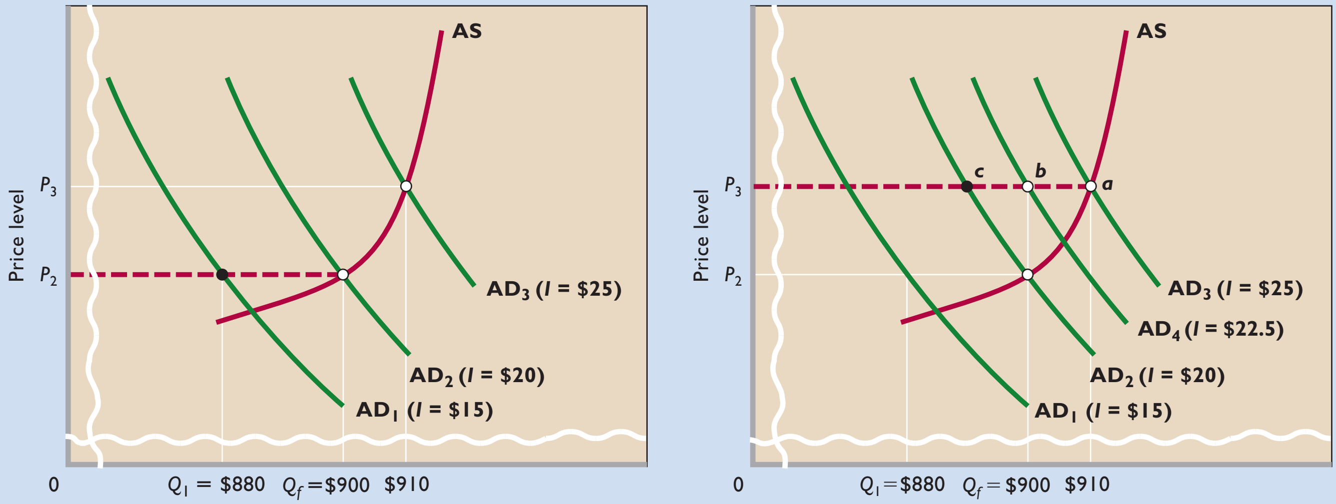
QUICK QUIZ FOR FIGURE 34.5

1. The ultimate objective of an expansionary monetary policy is depicted by:
   a. a decrease in the money supply from S_{m3} to S_{m2}.
   b. a reduction of the interest rate from 8 to 6 percent.
   c. an increase in investment from \$20 billion to \$25 billion.
   d. an increase in real GDP from Q_1 to Q_f.

investment spending on aggregate demand. (Ignore Figure 34.5d for the time being. We will return to it shortly.) As noted, aggregate demand curve AD_1 is associated with the \$15 billion level of investment, AD_2 with investment of \$20 billion, and AD_3 with investment of \$25 billion. That is, investment spending is one of the determinants of aggregate demand. Other things equal, the greater the investment spending, the farther to the right lies the aggregate demand curve.

Suppose the money supply in Figure 34.5a is \$150 billion (S_{m2}), producing an equilibrium interest rate of 8 percent. In Figure 34.5b we see that this 8 percent interest rate will bring forth \$20 billion of investment spending. This \$20 billion of investment spending joins with consumption spending, net exports, and government spending to yield aggregate demand curve AD_2 in Figure 34.5c. The equilibrium levels of real output and prices are Q_f = \$900 billion and P_2, as determined by the intersection of AD_2 and the aggregate supply curve AS.

To test your understanding of these relationships, explain why each of the other two levels of money supply in Figure 34.5a results in a different interest rate, level of investment, aggregate demand curve, and equilibrium real output.



Real domestic product, GDP
(billions of dollars)

(c)
Equilibrium real GDP
and the price level

Real domestic product, GDP
(billions of dollars)

(d)
Equilibrium real GDP
and the price level

2. A successful restrictive monetary policy is evidenced by a shift in the money supply curve from:
   a. S_{m3} to a point halfway between S_{m2} and S_{m3}, a decrease in investment from \$25 billion to \$22.5 billion, and a decline in aggregate demand from AD_3 to AD_4.
   b. S_{m1} to S_{m2}, an increase in investment from \$20 billion to \$25 billion, and an increase in real GDP from Q_1 to Q_f.
   c. S_{m3} to S_{m2}, a decrease in investment from \$25 billion to \$20 billion, and a decline in the price level from P_3 to P_2.
   d. S_{m3} to S_{m2}, a decrease in investment from \$25 billion to \$20 billion, and an increase in aggregate demand from AD_2 to AD_3.

3. The Federal Reserve could increase the money supply from S_{m1} to S_{m2} by:
   a. increasing the discount rate.
   b. reducing taxes.
   c. buying government securities in the open market.
   d. increasing the reserve requirement.

4. If the spending-income multiplier is 4 in the economy depicted, an increase in the money supply from \$125 billion to \$150 billion will:
   a. shift the aggregate demand curve rightward by \$20 billion.
   b. increase real GDP by \$25 billion.
   c. increase real GDP by \$100 billion.
   d. shift the aggregate demand curve leftward by \$5 billion.

Answers: 1. d; 2. a; 3. c; 4. a

Effects of an Expansionary Monetary Policy

Recall that the inflationary ratchet effect discussed in Chapter 30 describes the fact that real-world price levels tend to be downwardly inflexible. Thus, with our economy starting from the initial equilibrium where AD_2 intersects AS, the price level will be downwardly inflexible at P_2 so that aggregate supply will be horizontal to the left of Q_f. This means that if aggregate demand decreases, the economy's equilibrium will move leftward along the dashed horizontal line shown in Figure 34.5c.

Just such a decline would happen if the money supply fell to \$125 billion (S_{m1}), shifting the aggregate demand curve leftward to AD_1 in Figure 34.5c. This results in a real output of \$880 billion, \$20 billion less than the economy's full-employment output level of \$900 billion. The economy will be experiencing recession, a negative GDP gap, and substantial unemployment. The Fed therefore should institute an expansionary monetary policy.

To increase the money supply, the Fed will take some combination of the following actions: (1) buy government securities from banks and the public in the open market, (2) lower the legal reserve ratio, (3) lower the discount rate, and (4) reduce the interest rate that it pays on reserves. The intended outcome will be an increase in excess reserves in the commercial banking system and a decline in the federal funds rate. Because excess reserves are the basis on which commercial banks and thrifts can earn profit by lending and thus creating checkable-deposit money, the nation's money supply will rise. An increase in the money supply will lower the interest rate, increasing investment, aggregate demand, and equilibrium GDP.

For example, an increase in the money supply from \$125 billion to \$150 billion (S_{m1} to S_{m2}) will reduce the interest rate from 10 to 8 percent, as indicated in Figure 34.5a, and will boost investment from \$15 billion to \$20 billion, as shown in Figure 34.5b. This \$5 billion increase in investment will shift the aggregate demand curve rightward by more than the increase in investment because of the multiplier effect. If the economy's MPC is 0.75, the multiplier will be 4, meaning that the \$5 billion increase in investment will shift the AD curve rightward by \$20 billion (= 4 × \$5 billion) at each price level. Specifically, aggregate demand will shift from AD_1 to AD_2, as shown in Figure 34.5c. This rightward shift in the aggregate demand curve along the dashed horizontal line will eliminate the negative GDP gap by increasing GDP from \$880 billion to the full-employment GDP of Q_f = \$900 billion.[1]

Column 1 in Table 34.3 summarizes the chain of events associated with an expansionary monetary policy.

Effects of a Restrictive Monetary Policy

Next we consider restrictive monetary policy. To prevent overcrowding, we will use graphs *a*, *b*, and *d* (not *c*) in Figure 34.5 to demonstrate the effects of a restrictive monetary policy on the economy. Figure 34.5d represents exactly the same economy as Figure 34.5c but adds some extra curves that relate only to our explanation of restrictive monetary policy.

To see how restrictive monetary policy works, first consider a situation in which the economy moves from a full-employment equilibrium to operating at more than full employment so that inflation is a problem and restrictive monetary policy would be appropriate. Assume that the economy begins at the full-employment equilibrium where AD_2 and AS intersect. At this equilibrium, Q_f = \$900 billion and the price level is P_2.

Next, assume that the money supply expands from \$150 billion to \$175 billion (S_{m3}) in Figure 34.5a. This results in an interest rate of 6 percent, investment spending of \$25 billion rather than \$20 billion, and aggregate demand AD_3. As the AD curve shifts to the right from AD_2 to AD_3 in Figure 34.5d, the economy will move along the upsloping AS curve until it comes to an equilibrium at point *a*, where AD_3 intersects AS. At the new equilibrium, the price level has risen to P_3 and the equilibrium level of real GDP has risen to \$910 billion, indicating an inflationary GDP gap of \$10 billion (= \$910 billion − \$900 billion). Aggregate demand AD_3 is excessive relative to the economy's full-employment level of real output Q_f = \$900 billion. To rein in spending, the Fed will institute a restrictive monetary policy.

The Federal Reserve Board will direct Federal Reserve Banks to undertake some combination of the following actions: (1) sell government securities to banks and the public in the open market, (2) increase the legal reserve ratio, (3) increase the discount rate, and (4) increase the interest rate that it pays on reserves. Banks then will discover that their reserves are below those required and that the federal funds rate has increased. So they will need to reduce their checkable deposits by

TABLE 34.3 Monetary Policies for Recession and Inflation

(1) Expansionary Monetary Policy	(2) Restrictive Monetary Policy
Problem: unemployment and recession	*Problem:* inflation
↓	↓
Federal Reserve buys bonds, lowers reserve ratio, lowers the discount rate, or reduces the interest rate on reserves	Federal Reserve sells bonds, increases reserve ratio, raises the discount rate, or increases the interest rate on reserves
↓	↓
Excess reserves increase	Excess reserves decrease
↓	↓
Federal funds rate falls	Federal funds rate rises
↓	↓
Money supply rises	Money supply falls
↓	↓
Interest rate falls	Interest rate rises
↓	↓
Investment spending increases	Investment spending decreases
↓	↓
Aggregate demand increases	Aggregate demand decreases
↓	↓
Real GDP rises	Inflation declines

[1]To keep things simple, we assume that the increase in real GDP does not increase the demand for money. In reality, the transactions demand for money would rise, slightly dampening the decline in the interest rate shown in Figure 34.5a.

refraining from issuing new loans as old loans are paid back. This will shrink the money supply and increase the interest rate. The higher interest rate will discourage investment, lowering aggregate demand and restraining demand-pull inflation.

But the Fed must be careful about just how much to decrease the money supply. The problem is that the inflation ratchet will take effect at the new equilibrium point *a*, such that prices will be inflexible at price level P_3. As a result, the dashed horizontal line to the left of point *a* in Figure 34.5d will become relevant. This means that the Fed cannot simply lower the money supply to S_{m2} in Figure 34.5a. If it were to do that, investment demand would fall to \$20 billion in Figure 34.5b and the AD curve would shift to the left from AD_3 back to AD_2. But because of inflexible prices, the economy's equilibrium would move to point *c*, where AD_2 intersects the horizontal dashed line to the left of point *a*. This would put the economy into a recession, with equilibrium output below the full-employment output level of Q_f = \$900 billion.

What the Fed needs to do to achieve full employment is to move the AD curve back only from AD_3 to AD_4, so that the economy will come to equilibrium at point *b*. This will require a \$10 billion decrease in aggregate demand, so that equilibrium output falls from \$910 billion at point *a* to Q_f = \$900 billion at point *b*. The Fed can achieve this shift by setting the supply of money in Figure 34.5a at \$162.5 billion. To see how this works, draw in a vertical money supply curve in Figure 34.5a at \$162.5 billion and label it as S_{m4}. It will be exactly halfway between money supply curves S_{m2} and S_{m3}. Notice that the intersection of S_{m4} with the money demand curve D_m will result in an interest rate of 7 percent. In Figure 34.5b, this interest rate of 7 percent will result in investment spending of \$22.5 billion (halfway between \$20 billion and \$25 billion). Thus, by setting the money supply at \$162.5 billion, the Fed can reduce investment spending by \$2.5 billion, lowering it from the \$25 billion associated with AD_3 down to only \$22.5 billion. This decline in investment spending will initially shift the AD curve only \$2.5 billion to the left of AD_3. But then the multiplier process will work its magic. Since the multiplier is 4 in our model, the AD curve will end up moving by a full \$10 billion (= 4 × \$2.5 billion) to the left, to AD_4. This shift will move the economy to equilibrium *b*, returning output to the full employment level and eliminating the inflationary GDP gap.[2]

[2]Again, we assume for simplicity that the decrease in nominal GDP does not feed back to reduce the demand for money and thus the interest rate. In reality, this would occur, slightly dampening the increase in the interest rate shown in Figure 34.5a.

Column 2 in Table 34.3 summarizes the cause-effect chain of a tight money policy.

QUICK REVIEW 34.5

- The Fed is engaging in an expansionary monetary policy when it increases the money supply to reduce interest rates and increase investment spending and real GDP.
- The Fed is engaging in a restrictive monetary policy when it reduces the money supply to increase interest rates and reduce investment spending and inflation.

Monetary Policy: Evaluation and Issues

LO34.6 Explain the effectiveness of monetary policy and its shortcomings.

Monetary policy has become the dominant component of U.S. national stabilization policy. It has two key advantages over fiscal policy:

- Speed and flexibility.
- Isolation from political pressure.

Compared with fiscal policy, monetary policy can be quickly altered. Recall that congressional deliberations may delay the application of fiscal policy for months. In contrast, the Fed can buy or sell securities from day to day and thus affect the money supply and interest rates almost immediately.

Also, because members of the Fed's Board of Governors are appointed and serve 14-year terms, they are relatively isolated from lobbying and need not worry about retaining their popularity with voters. Thus, the Board, more readily than Congress, can engage in politically unpopular policies (higher interest rates) that may be necessary for the long-term health of the economy. Moreover, monetary policy is a subtler and more politically neutral measure than fiscal policy. Changes in government spending directly affect the allocation of resources, and changes in taxes can have extensive political ramifications. Because monetary policy works more subtly, it is more politically palatable.

Recent U.S. Monetary Policy

The Fed has been highly active in its use of monetary policy in recent decades.

The 2001 Recession In 2000, the economy abruptly slowed after a long period of full employment and strong economic growth. The Fed responded to the slowdown by cutting the federal funds interest rate by a full percentage point in two increments in January 2001. Despite those rate cuts, the economy entered a recession in March 2001. Between March 20, 2001, and August 21, 2001, the Fed reduced the federal funds rate from 5 percent to 3.5 percent in a series of steps. In the 3 months following the terrorist attacks of September 11, 2001, it lowered the federal funds rate from 3.5 percent to 1.75 percent, and it left the rate there until it lowered it to 1.25 percent in November 2002. Partly because of the Fed's actions, the prime interest rate dropped from 9.5 percent at the end of 2000 to 4.25 percent in December 2002.

Economists generally give the Fed high marks for helping to keep the recession of 2001 relatively mild, particularly in view of the adverse economic impacts of the terrorist attacks of September 11, 2001, and the steep stock-market decline in 2001–2002.

The Fed left the federal funds rate at historic lows in 2003. But as the economy began to expand robustly in 2004, the Fed engineered a gradual series of rate hikes designed to boost the prime interest rate and other interest rates to make sure that aggregate demand continued to grow at a pace consistent with low inflation. By the summer of 2006, the target for the federal funds rate had risen to 5.25 percent and the prime rate was 8.25 percent. With the economy enjoying sustainable, noninflationary growth, the Fed left the federal funds rate at 5.25 percent for over a year.

The 2007–2009 Recession The mortgage default crisis (discussed in Chapter 32) began during the late summer of 2007 and posed a grave threat to the financial system and the economy. In response, the Fed took several actions. In August it lowered the discount rate by half a percentage point. Then, between September 2007 and April 2008, it lowered the target for the federal funds rate from 5.25 percent to 2 percent. And as discussed in Chapter 32, the Fed also took a series of extraordinary actions to prevent the failure of key financial firms.

In October 2008, the Fed first reduced the federal funds target rate to 1.5 percent and then later that same month to 1 percent. In December 2008, the Fed lowered it further to a targeted range of 0 percent to 0.25 percent. That near-zero targeted range was by far the lowest in history. Viewed through Figure 34.3, the Fed aggressively pushed the supply of federal funds curve downward (increased the supply of federal funds) to lower the actual federal funds rate to its target level. All these monetary actions and lender-of-last-resort functions helped to stabilize the banking sector and keep credit flowing—thereby offsetting at least some of the damage done by the financial crisis.

The decline in the federal funds rate to near zero during the financial and economic crisis dropped the prime interest rate (review Figure 34.4). In December 2007, the prime interest rate stood at 7.3 percent. By January 2009, it had declined to 3.25 percent, where it remained through 2013.

The Federal Reserve is lauded by most observers for its quick and innovative actions during the financial crisis and severe recession. Nevertheless, some economists contend that the Fed contributed to the financial crisis by holding the federal funds interest rate too low for too long during the recovery from the 2001 recession. These critics say that the artificially low interest rates made mortgage and other loans too inexpensive and therefore contributed to the borrowing frenzy by homeowners and other financial investors. Other economists counter that the low mortgage interest rates resulted from huge inflows of savings from abroad to a wide variety of U.S. financial markets.

After the Great Recession The U.S. economy recovered very slowly from the Great Recession, especially in terms of employment. After falling from a peak of 138.1 million in January 2008 to a trough of 129.9 million in September 2010, the total number of people with jobs rebounded to just 135.5 million by April 2013. Thus, nearly four years after the recession officially ended in the summer of 2009, 2.6 million fewer people were employed than before the recession. By contrast, after all other post–World War II recessions, employment had fully recovered within four years—and in most cases within two years.

The Federal Reserve understood the depth of the economy's problems early on and responded with a series of innovative monetary policy initiatives designed to stimulate GDP and employment growth.

Zero Interest Rate Policy The Fed began by moving toward a **zero interest rate policy,** or ZIRP, in December 2008. Under ZIRP, the Fed aimed to keep short-term interest rates near zero to stimulate the economy. To that end, open-market operations were used to keep the federal funds rate between zero and 0.25 percent.

Zero Lower Bound Problem After ZIRP was implemented and interest rates were pushed toward zero, the economic growth remained weak. That implied that the Fed would have to figure out a way to deal with the **zero lower bound problem,** under which a central bank is constrained in its ability to stimulate the economy through lower interest rates by the fact that nominal interest rates cannot be driven lower than zero.

Why can't nominal interest rates be driven lower than zero? Because if nominal interest rates were negative, people would not want to put their money into banks because doing so would mean that their balances would shrink over time (rather than grow over time, as they do when interest rates are positive). Thus, any central bank that attempted to impose negative nominal interest rates would see deposits withdrawn from banks. That could be economically catastrophic because if people withdrew deposits from banks, banks would have much less money to lend out to consumers and entrepreneurs. The monetary multiplier of Chapter 33 would work in reverse and the supply of lending and credit in the economy would decrease precipitously, thereby negatively affecting aggregate demand.

Quantitative Easing The Fed's response to the zero lower bound problem was **quantitative easing,** or QE. In terms of mechanics, quantitative easing looks exactly like open-market operations, with the Fed purchasing bonds in order increase the amount of reserves in the banking system. QE differs from ordinary open-market operations, however, in that it is not intended to lower interest rates. Under QE, the Fed buys bonds solely with the intention of increasing the quantity of reserves in the banking system. Interest rates remain at the same levels to avoid the zero lower bound problem. But with bank reserves increased, the economy will hopefully be stimulated through increased lending.

Another difference between QE and regular open-market operations is that QE can involve the purchase of not only U.S. government bonds but also debt issued by government agencies or government-backed corporations (which are known as government-sponsored entities, or GSEs).

The first round of quantitative easing began in March 2009 and involved the Fed purchasing \$1.75 trillion worth of bonds. The bonds consisted of \$300 billion worth of Treasury bonds and \$1.45 trillion worth of bonds issued by either U.S. government agencies or the two government-backed mortgage lenders, Freddie Mac and Fannie Mae.

The second round of quantitative easing ("QE2") began in November 2010 and involved the Fed telling the public of its intention to purchase \$600 billion of U.S. Treasury bonds at the rate of \$75 billion per month over the following eight months.

Forward Commitment The innovative feature of QE2 was that the Fed engaged in **forward commitment,** preannouncing exactly how much it was going to buy during QE2 and how long the buying would last. This was a major change in monetary policy because up to that time the Fed had (along with most central banks) stuck to a policy of being vague about how long any particular policy initiative would last. For instance, if the Fed lowered or raised the federal funds target rate, it did not publicly announce for how long the change would last.

The rationale behind being vague was to preserve the Fed's flexibility to make changes if unexpected circumstances arose. However, that flexibility came at the cost of reduced credibility because the public might not react strongly to a policy change if it believed that the policy change might be reversed at any moment. By preannouncing the exact size and duration of QE, the Fed removed that worry. By making forward commitments, the public would know not only the content of a policy change but also that it wasn't going to be suddenly reversed.

With respect to the banking system, the announcement of both the size and duration of the Fed's open-market purchases of bonds meant that banks would know that the resulting increases in reserves would not suddenly be reversed. That would make the banks more likely to lend those new reserves because they wouldn't have any nagging doubts that the Fed might suddenly reverse policy, reduce reserves, and force the banks to suddenly and unexpectedly reduce their lending activities.

Operation Twist The Fed's use of forward commitment continued in September 2011, when it began the Maturity Extension Program, commonly known as Operation Twist. Under that program, the Fed preannounced that, by the end of 2012, it would purchase \$677 billion in long-term government bonds while simultaneously selling an equivalent dollar amount of short-term government bonds. The Fed's motivation for doing so was to spur investment and consumption by reducing long-term interest rates, which were at that time several percentage points higher than short-term interest rates (which remained near zero thanks to ZIRP).

The intended reduction in long-term interest rates was accomplished by purchasing long-term bonds and driving up their prices. The money needed for those purchases was provided by selling an equivalent dollar amount of short-term bonds. Crucially, the amount of short-term bonds sold by the Fed was not nearly enough to alter short-term interest rates. Thus, they stayed near zero while longer-term rates fell.

QE3 When economic growth remained weak in 2012, the Fed decided that the lack of effectiveness of QE2 and

Operation Twist might have been due to the fact that both had featured limited time durations. Thus, banks might have been worried that the Fed might reverse policy as soon as the limited time durations of QE2 and Operation Twist came to an end.

To avoid that possibility going forward, the Fed's announcement of QE3 in September 2012 involved explicitly stating not only that the Fed would purchase $85 billion per month in bonds, but that those purchases had no specific end date and would in fact continue until the employment situation improved substantially. By making an open-ended commitment, the Fed hoped to enhance the credibility of its monetary stimulus policies.

In September 2012, the Fed also issued an open-ended policy commitment with respect to the federal funds rate. The Fed announced that the federal funds target rate would remain "exceptionally low" as long as the unemployment rate stayed above 6.5 percent and inflation remained muted, at 2 percent per year or less. Thus, the Fed committed itself to utilizing ZIRP until either the jobs situation dramatically improved or inflation started rising too high.

That forward commitment allowed businesses, banks, and consumers to have a much better sense of how long monetary stimulus would last and the circumstances under which it would be cut off. As the economy moved into 2013, the Fed hoped that its increasingly specific forward guidance would help to improve the effectiveness of its monetary stimulus efforts.

Problems and Complications

Despite its recent successes in the United States, monetary policy has certain limitations and faces actual-economy complications.

Lags Recall that fiscal policy is hindered by three delays, or lags—a recognition lag, an administrative lag, and an operational lag. Monetary policy also faces a recognition lag and an operational lag, but because the Fed can decide and implement policy changes within days, it avoids the long administrative lag that hinders fiscal policy.

A recognition lag affects monetary policy because normal monthly variations in economic activity and the price level mean that the Fed may not be able to quickly recognize when the economy is truly starting to recede or when inflation is really starting to rise. Once the Fed acts, an operation lag of 3 to 6 months affects monetary policy because that much time is typically required for interest-rate changes to have their full impacts on investment, aggregate demand, real GDP, and the price level. These two lags complicate the timing of monetary policy.

CONSIDER THIS . . .

Up, Up, and Away

The consolidated balance sheet of the 12 Federal Reserve Banks changed markedly during the severe recession of 2007–2009. Total Fed assets increased from $885 billion in February 2008 to $2,317 billion in March 2010. This increase reflected an enormous rise in the number of U.S. securities, mortgage-backed securities, and other financial assets purchased by the Federal Reserve. In undertaking its monetary policy and its lender-of-last-resort functions, the Fed bought these securities from financial institutions—purposely increasing the liquidity of the financial system.

On the liability side, the reserves of commercial banks rose from $43 billion in February 2008 to $1,148 billion in March 2010. To make sure they were liquid and the funds were safe, banks placed much of the proceeds from selling securities to the Fed into their respective reserve accounts at the Fed. This flow was strengthened because the Fed began paying interest on the reserves that banks were holding at the Fed.

In March 2010 total bank reserves held at the Fed exceeded total checkable deposits held by the banks. The severe distress in the financial system had voluntarily turned the fractional reserve system into a 100-percent-plus reserve system! The banks had enormous excess reserves from which to increase lending once the banks became more certain of their own financial viability and the likelihood that newly issued loans would be paid back.

The Fed's use of quantitative easing caused the Fed's balance sheet to increase even further after the Great Recession ended. By May 2013, it had reached $3.3 trillion and was continuing to grow by $85 billion per month as QE3 continued. The Fed therefore faces the challenging task of using monetary policy to absorb large portions of this overstock of excess reserves as the economy recovers and picks up momentum. It does not want the banks to lend out the full amount of these excess reserves because that would flood the economy with bank-created money and excessively expand the money supply. During a vigorous economic expansion, the excessive money and resulting very low interest rates could produce such large expansions of aggregate demand that rapid inflation would occur.

Cyclical Asymmetry and the Liquidity Trap Monetary policy may be highly effective in slowing expansions and controlling inflation but may be much less reliable in pushing the economy from a severe recession. Economists say that monetary policy may suffer from **cyclical asymmetry.** The metaphor of "pushing on a string" is often invoked to capture this problem. Imagine the Fed standing on the left-hand side of Figure 34.5d, holding one end of a "monetary-policy string." And imagine that the other end of the monetary-policy string is tied to the AD curve. Because the string would go taut if pulled on, monetary policy may be useful in *pulling* aggregate demand to the left. But because the string would go limp if pushed on, monetary policy will be rather ineffective at *pushing* aggregate demand to the right.

The reason for this asymmetry has to do with the asymmetric way in which people may act in response to changes in bank reserves. If pursued vigorously, a restrictive monetary policy can deplete commercial banking reserves to the point where banks are forced to reduce the volume of loans. That means a contraction of the money supply, higher interest rates, and reduced aggregate demand. The Fed can absorb sufficient reserves and eventually achieve its goal.

But the Fed cannot be certain of achieving its goal when it adds reserves to the banking system because of the so-called **liquidity trap,** in which adding more liquidity to banks has little or no additional positive effect on lending, borrowing, investment, or aggregate demand. For example, during the recent recession, the Fed created billions of dollars of excess reserves that drove down the federal funds rate to as low as 0.2 percent. The prime interest rate fell from 7.3 percent (December 2007) to 3.25 percent (March 2009). Nevertheless, lending by banks stalled throughout the first 15 months of the recession and remained weak even after the Fed implemented ZIRP, QE, Operation Twist, and forward commitments over the following four years. The banks were fearful that the loans they would make to households, businesses, and other financial institutions would not be paid back. Consequently, they were content to hold reserves at the Federal Reserve Banks.

To switch analogies, an expansionary monetary policy can suffer from a "you can lead a horse to water, but you can't make it drink" problem. The Fed can create excess reserves, but it cannot guarantee that the banks will actually make additional loans and thus promote spending. If commercial banks seek liquidity and are unwilling to lend, the efforts of the Fed will be of little avail. Similarly, households and businesses can frustrate the intentions of the Fed by not borrowing excess reserves being made available as loans. And when the Fed buys securities from the public, people may choose to pay off existing loans with the money received, rather than increasing their spending on goods and services.

Furthermore, a severe recession may so undermine business confidence that the investment demand curve shifts to the left and overwhelms the lower interest rates associated with an expansionary monetary policy. That is what happened in the most recent recession. Although the Fed drove the real interest rate down to zero percent, investment spending remained low and the economy remained mired in recession. The recent U.S. experience reminds us that active monetary policy certainly is not a cure-all for the business cycle. Under some circumstances, monetary policy may be like "pushing on a string."

The liquidity trap that occurred during the severe recession was a primary reason why public policy in the United States turned so significantly and forcefully toward fiscal policy in 2009. Recall our discussion of the American Recovery and Redevelopment Act of 2009, which authorized the infusion of $787 billion of new tax cuts and government spending in 2009 and 2010.

QUICK REVIEW 34.6

- The Fed aggressively lowered the federal funds interest rate following 9/11 and the 2001 recession and also during the severe recession of 2007–2009.
- To help stimulate the economy after the Great Recession, the Fed implemented the zero interest rate policy (ZIRP), quantitative easing (QE), Operation Twist, and forward commitment.
- The main strengths of monetary policy are (a) speed and flexibility and (b) political acceptability; its main weaknesses are (a) time lags and (b) potential ineffectiveness during severe recession.

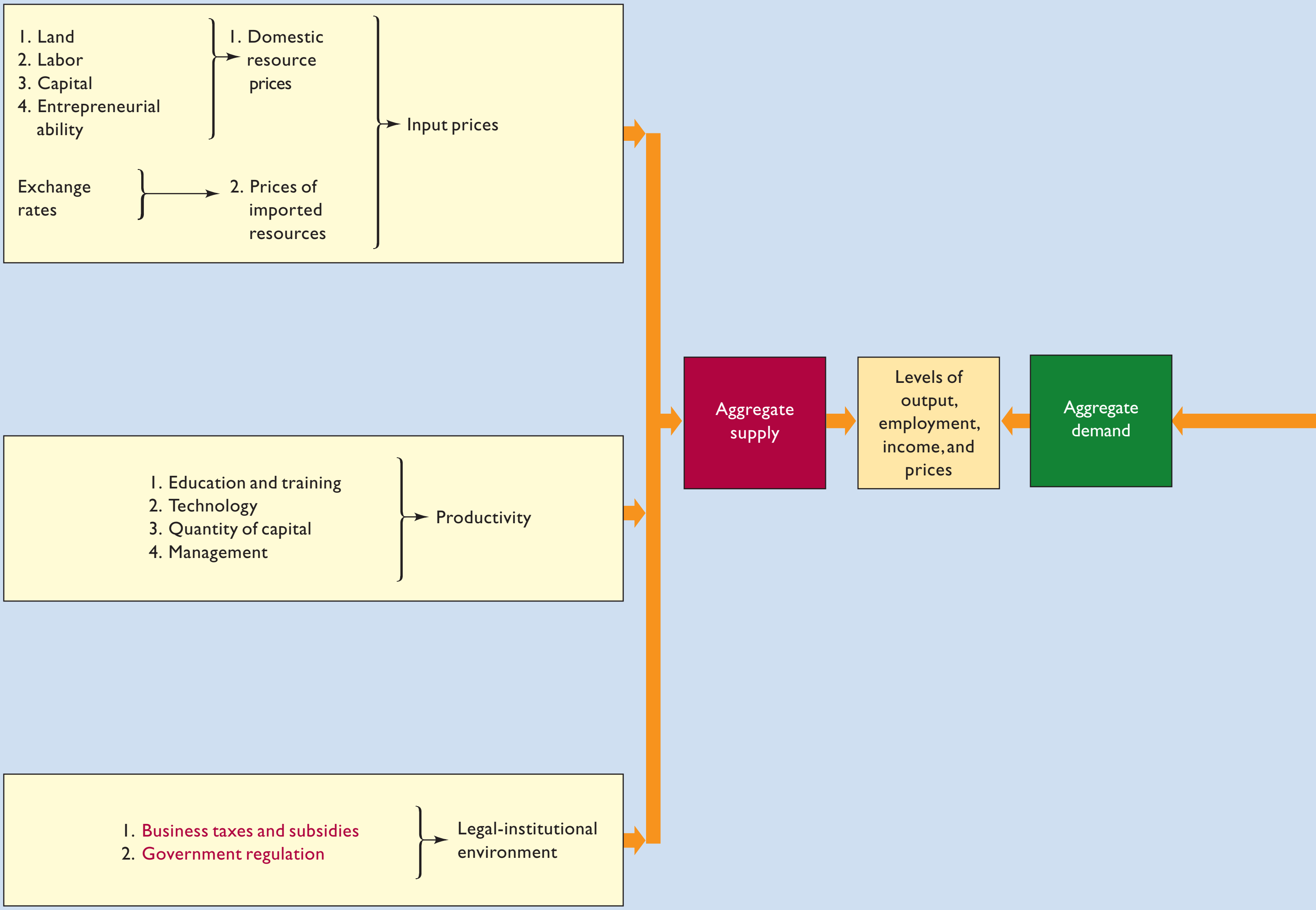
The "Big Picture"

Figure 34.6 (Key Graph) on pages 770 and 771 brings together the analytical and policy aspects of macroeconomics discussed in this and the eight preceding chapters. This "big picture" shows how the many concepts and principles discussed relate to one another and how they constitute a coherent theory of the price level and real output in a market economy.

Study this diagram and you will see that the levels of output, employment, income, and prices all result from the interaction of aggregate supply and aggregate demand. The items shown in red relate to public policy.

KEY GRAPH

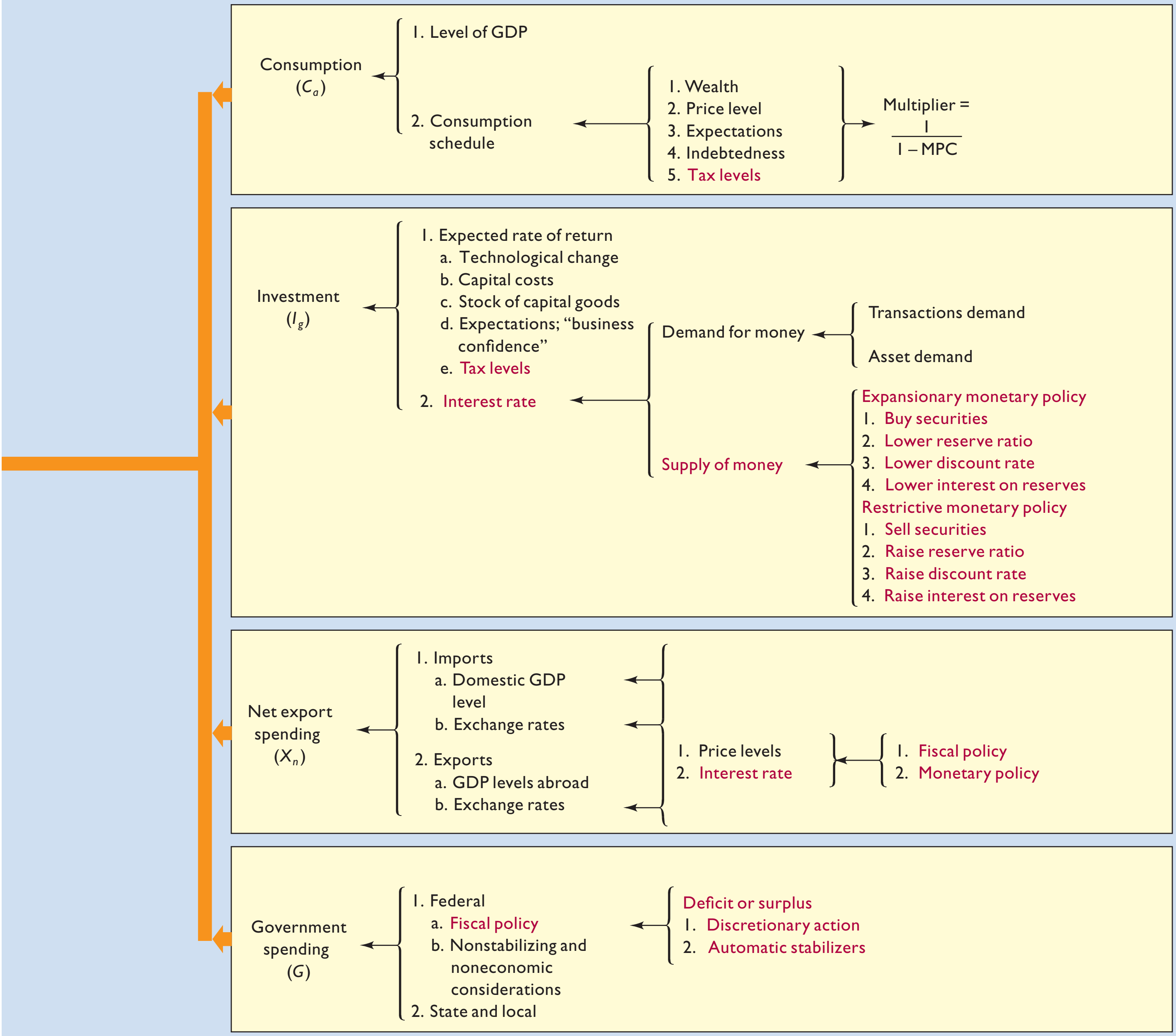
FIGURE 34.6 The AD-AS theory of the price level, real output, and stabilization policy. This figure integrates the various components of macroeconomic theory and stabilization policy. Determinants that either constitute public policy or are strongly influenced by public policy are shown in red.



QUICK QUIZ FOR FIGURE 34.6

1. All else equal, an increase in domestic resource availability will:
   a. increase input prices, reduce aggregate supply, and increase real output.
   b. raise labor productivity, reduce interest rates, and lower the international value of the dollar.
   c. increase net exports, increase investment, and reduce aggregate demand.
   d. reduce input prices, increase aggregate supply, and increase real output.

2. All else equal, an expansionary monetary policy during a recession will:
   a. lower the interest rate, increase investment, and reduce net exports.
   b. lower the interest rate, increase investment, and increase aggregate demand.
   c. increase the interest rate, increase investment, and reduce net exports.
   d. reduce productivity, aggregate supply, and real output.



3. A personal income tax cut, combined with a reduction in corporate income and excise taxes, would:

a. increase consumption, investment, aggregate demand, and aggregate supply.
b. reduce productivity, raise input prices, and reduce aggregate supply.
c. increase government spending, reduce net exports, and increase aggregate demand.
d. increase the supply of money, reduce interest rates, increase investment, and expand real output.

4. An appreciation of the dollar would:

a. reduce the price of imported resources, lower input prices, and increase aggregate supply.
b. increase net exports and aggregate demand.
c. increase aggregate supply and aggregate demand.
d. reduce consumption, investment, net export spending, and government spending.

Answers: 1. d; 2. b; 3. a; 4. a

LAST WORD

Worries about ZIRP, QE, and Twist

ZIRP, QE, and Operation Twist Provided Massive Economic Stimulus During and After the Great Recession. But There Remain Many Worries About Unintended Consequences.

When the financial crisis reached its peak in 2008, the Fed acted aggressively to prevent bank runs and stabilize the financial system by acting as a lender of last resort. It also did its best to get the economy moving again by lowering short-term interest rates to nearly zero—a strategy that came to be known as the zero interest rate policy, or ZIRP.

When ZIRP by itself didn't seem to be causing enough stimulus, the Fed also began engaging in trillions of dollars' worth of bond purchases. Those purchases went by the name of quantitative easing, or QE, because the Fed printed up electronic money to pay for the purchases, thereby massively increasing (easing) the total quantity of money in circulation. The Fed's hope was that the additional money would lead to additional spending and lending that would boost aggregate demand by increasing consumption and investment. Later, the policy known as Operation Twist lowered longer-term interest rates.

One important effect of ZIRP and Operation Twist was to help the U.S. federal government engage in aggressive deficit-financed fiscal stimulus. Thanks to ZIRP and Operation Twist, the federal government was able to fund large deficits by issuing 10-year bonds at nominal interest rates of about 2 percent—substantially lower than the historical average of about 6 percent.

But the Federal Reserve didn't just help the federal government with low

SUMMARY

LO34.1 Discuss how the equilibrium interest rate is determined in the market for money.

The total demand for money consists of the transactions demand for money plus the asset demand for money. The amount of money demanded for transactions varies directly with the nominal GDP; the amount of money demanded as an asset varies inversely with the interest rate. The market for money combines the total demand for money with the money supply to determine equilibrium interest rates.

Interest rates and bond prices are inversely related.

LO34.2 Describe the balance sheet of the Federal Reserve and the meaning of its major items.

The consolidated balance sheet of the Federal Reserve System lists the collective assets and liabilities of the 12 Federal Reserve banks. The assets consist largely of Treasury notes, Treasury bills, and Treasury bonds. The major liabilities are reserves of commercial banks, Treasury deposits, and Federal Reserve notes outstanding. The balance sheet is useful in understanding monetary policy because open-market operations increase or decrease the Fed's assets and liabilities.

LO34.3 List and explain the goals and tools of monetary policy.

The goal of monetary policy is to help the economy achieve price stability, full employment, and economic growth.

The four main instruments of monetary policy are (*a*) open-market operations, (*b*) the reserve ratio, (*c*) the discount rate, and (*d*) interest on reserves.

LO34.4 Describe the federal funds rate and how the Fed directly influences it.

The federal funds rate is the interest rate that banks charge one another for overnight loans of reserves. The prime interest rate is the interest rates. It also served as the federal government's primary lender. In 2012, for instance, the Federal Reserve purchased over 70 percent of all U.S. government debt. Thus, 70 percent of the federal government's new borrowing came from the Federal Reserve in the form of newly printed money that the Fed created in order to fund its open-market purchases of government bonds.

The consensus among economists was that the Fed's aggressive use of ZIRP and QE were warranted by the severity of the financial crisis and the historically slow pace with which the economy recovered after the 2007–2009 recession. However, concerns were also raised about possible unintended consequences.

One worry had to do with the large annual budget deficits that the federal government was running. While many economists felt that the large deficits were appropriate given the sluggish economy, others believed that the federal government was overspending and taking resources away from the private sector. As a result, they felt that the Fed's use of ZIRP and QE was making it too easy for Congress to overspend and run large budget deficits because the Fed would always provide a ready buyer for the bonds that had to be issued to finance those large deficits.

A longer-term worry was that when ZIRP ended and interest rates began to rise again toward normal levels, the federal government would be suddenly confronted with huge interest costs. Consider the $16 trillion of debt that had accumulated by 2013: $16 trillion borrowed at 2 percent interest generates annual interest payments of $320 billion per year. But if the interest rate on government debt were to rise back to its historical average of 6 percent, the annual interest payments on $16 trillion would come to $960 billion per year. Such a huge increase in annual interest payments would likely require either massive budget cuts or even more borrowing, unless the economy began to grow so quickly that increased tax revenues were enough to compensate for the increased interest payments.

Another problem with extremely low interests rates is that they punish savers. A senior citizen who has saved for retirement will find that her investments yield very low rates of return when the Fed is keeping interest rates low. Instead of being able to live off of the interest generated by her investments, she may find herself spending down her accumulated savings because the interest payments amount to nearly nothing.

On a larger scale, pension plans and retirement funds are also hit hard by low interest rates. Those institutions take deposits from current workers, invest those funds, and promise to pay out certain amounts when workers retire. Prior to the financial crisis, most of those institutions had assumed that they would be able to earn 8 percent per year on the retirement funds that they were entrusted with. But with the Fed keeping interest rates so low, the pension plans and retirement funds were not earning anywhere near 8 percent per year on their investments. As a result, the low interest rates engineered by the Fed made it very unlikely that pension plans and retirement funds would be able to keep their promises to retirees and deliver enough money in 20 or 30 years to pay each individual retiree what he or she had been promised.

benchmark rate that banks use as a reference rate for a wide range of interest rates on short-term loans to businesses and individuals.

The Fed adjusts the federal funds rate to a level appropriate for economic conditions. Under an expansionary monetary policy, it purchases securities from commercial banks and the general public to inject reserves into the banking system. This lowers the federal funds rate to the targeted level and also reduces other interest rates (such as the prime rate). Under a restrictive monetary policy, the Fed sells securities to commercial banks and the general public via open-market operations. Consequently, reserves are removed from the banking system, and the federal funds rate and other interest rates rise.

LO34.5 Identify the mechanisms by which monetary policy affects GDP and the price level.

Monetary policy affects the economy through a complex cause-effect chain: (*a*) policy decisions affect commercial bank reserves; (*b*) changes in reserves affect the money supply; (*c*) changes in the money supply alter the interest rate; (*d*) changes in the interest rate affect investment; (*e*) changes in investment affect aggregate demand; (*f*) changes in aggregate demand affect the equilibrium real GDP and the price level. Table 34.3 draws together all the basic ideas relevant to the use of monetary policy.

LO34.6 Explain the effectiveness of monetary policy and its shortcomings.

The advantages of monetary policy include its flexibility and political acceptability. In recent years, the Fed has used monetary policy to keep inflation low while helping limit the depth of the recession of 2001, to boost the economy as it recovered from that recession, to help stabilize the banking sector in the wake of the mortgage debt crisis, and to promote recovery from the severe recession of 2007–2009. Today, nearly all economists view monetary policy as a significant stabilization tool.

Monetary policy has two major limitations and potential problems: (*a*) recognition and operation lags complicate the timing of monetary policy; (*b*) in a severe recession, the reluctance of banks to lend excess reserves and firms to borrow money to spend on capital goods may contribute to a liquidity trap that limits the effectiveness of an expansionary monetary policy.

TERMS AND CONCEPTS

monetary policy
interest
transactions demand for money
asset demand for money
total demand for money
open-market operations
reserve ratio
discount rate
interest on reserves
federal funds rate
expansionary monetary policy
prime interest rate
restrictive monetary policy
Taylor rule
zero interest rate policy (ZIRP)
zero lower bound problem
quantitative easing (QE)
forward commitment
cyclical asymmetry
liquidity trap

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. What is the basic determinant of (*a*) the transactions demand and (*b*) the asset demand for money? Explain how these two demands can be combined graphically to determine total money demand. How is the equilibrium interest rate in the money market determined? Use a graph to show the effect of an increase in the total demand for money on the equilibrium interest rate (no change in money supply). Use your general knowledge of equilibrium prices to explain why the previous interest rate is no longer sustainable. **LO34.1**
2. What is the basic objective of monetary policy? What are the major strengths of monetary policy? Why is monetary policy easier to conduct than fiscal policy? **LO34.3**
3. Distinguish between the federal funds rate and the prime interest rate. Why is one higher than the other? Why do changes in the two rates closely track one another? **LO34.4**
4. Why is a decrease in the supply of federal funds shown as an upshift of the supply curve in Figure 34.3, whereas an increase in federal funds is shown as a downshift of the supply curve? **LO34.4**
5. Suppose that you are a member of the Board of Governors of the Federal Reserve System. The economy is experiencing a sharp rise in the inflation rate. What change in the federal funds rate would you recommend? How would your recommended change get accomplished? What impact would the actions have on the lending ability of the banking system, the real interest rate, investment spending, aggregate demand, and inflation? **LO34.5**
6. Explain the links between changes in the nation's money supply, the interest rate, investment spending, aggregate demand, real GDP, and the price level. **LO34.5**
7. What do economists mean when they say that monetary policy can exhibit cyclical asymmetry? How does the idea of a liquidity trap relate to cyclical asymmetry? Why is this possibility of a liquidity trap significant to policymakers? **LO34.6**
8. **LAST WORD** Did Operation Twist target long-term or short-term interest rates? How does ZIRP cause problems for savers and pension funds? How might low interest rates lead to problematic fiscal policy decisions?

REVIEW QUESTIONS

1. When bond prices go up, interest rates go ___________. **LO34.1**
   a. Up.
   b. Down.
   c. Nowhere.
2. A commercial bank sells a Treasury bond to the Federal Reserve for $100,000. The money supply: **LO34.3**
   a. Increases by $100,000.
   b. Decreases by $100,000.
   c. Is unaffected by the transaction.
3. Use commercial bank and Federal Reserve Bank balance sheets to demonstrate the effect of each of the following transactions on commercial bank reserves: **LO34.3**
   a. Federal Reserve Banks purchase securities from banks.
   b. Commercial banks borrow from Federal Reserve Banks at the discount rate.
   c. The Fed reduces the reserve ratio.
   d. Commercial banks increase their reserves after the Fed increases the interest rate that it pays on reserves.
4. A bank currently has $100,000 in checkable deposits and $15,000 in actual reserves. If the reserve ratio is 20 percent, the bank has ___________ in money-creating potential. If the reserve ratio is 14 percent, the bank has ___________ in money-creating potential. **LO34.3**
   a. $20,000; $14,000.
   b. $3,000; $2,100.
   c. −$5,000; $1,000.
   d. $5,000; $1,000.

5. A bank borrows $100,000 from the Fed, leaving a $100,000 Treasury bond on deposit with the Fed to serve as collateral for the loan. The discount rate that applies to the loan is 4 percent and the Fed is currently mandating a reserve ratio of 10 percent. How much of the $100,000 borrowed by the bank must it keep as required reserves? **LO34.3**
   a. $0.
   b. $4,000.
   c. $10,000.
   d. $100,000.
6. Which of the following Fed actions will increase bank lending? **LO34.3**
   Select ***one or more*** *answers from the choices shown.*
   a. The Fed raises the discount rate from 5 percent to 6 percent.
   b. The Fed raises the reserve ratio from 10 percent to 11 percent.
   c. The Fed buys $400 million worth of Treasury bonds from commercial banks.
   d. The Fed lowers the discount rate from 4 percent to 2 percent.
7. If the Federal Reserve wants to increase the federal funds rate using open-market operations, it should __________ bonds. **LO34.4**
   a. Buy.
   b. Sell.
8. True or False: A liquidity trap occurs when expansionary monetary policy fails to work because an increase in bank reserves by the Fed does not lead to an increase in bank lending. **LO34.6**
9. True or False: In the United States, monetary policy has two key advantages over fiscal policy: (1) isolation from political pressure and (2) speed and flexibility. **LO34.6**

PROBLEMS

1. Assume that the following data characterize the hypothetical economy of Trance: money supply = $200 billion; quantity of money demanded for transactions = $150 billion; quantity of money demanded as an asset = $10 billion at 12 percent interest, increasing by $10 billion for each 2-percentage-point fall in the interest rate. **LO34.1**
   a. What is the equilibrium interest rate in Trance?
   b. At the equilibrium interest rate, what are the quantity of money supplied, the total quantity of money demanded, the amount of money demanded for transactions, and the amount of money demanded as an asset in Trance?
2. Suppose a bond with no expiration date has a face value of $10,000 and annually pays a fixed amount of interest of $800. In the table provided to the right, calculate and enter either the interest rate that the bond would yield to a bond buyer at each of the bond prices listed or the bond price at each of the interest yields shown. What generalization can be drawn from the completed table? **LO34.1**

Bond Price	Interest Yield, %
$ 8,000	______
______	8.9
$10,000	______
$11,000	______
______	6.2

3. In the tables that follow you will find consolidated balance sheets for the commercial banking system and the 12 Federal Reserve Banks. Use columns 1 through 3 to indicate how the balance sheets would read after each of transactions *a* to *c* is completed. Do not cumulate your answers; that is, analyze each transaction separately, starting in each case from the numbers provided. All accounts are in billions of dollars. **LO34.3**
   a. A decline in the discount rate prompts commercial banks to borrow an additional $1 billion from the Federal Reserve Banks. Show the new balance-sheet numbers in column 1 of each table.

		Consolidated Balance Sheet: All Commercial Banks		
		(1)	(2)	(3)
Assets:				
Reserves	$33	______	______	______
Securities	60	______	______	______
Loans	60	______	______	______
Liabilities and net worth:				
Checkable deposits	$150	______	______	______
Loans from the Federal Reserve Banks	3	______	______	______

		Consolidated Balance Sheet: The 12 Federal Reserve Banks		
		(1)	(2)	(3)
Assets:				
Securities	$60			
Loans to commercial banks.	3			
Liabilities and net worth:				
Reserves of commercial banks	$33			
Treasury deposits.	3			
Federal Reserve Notes	27			

b. The Federal Reserve Banks sell $3 billion in securities to members of the public, who pay for the bonds with checks. Show the new balance-sheet numbers in column 2 of each table.

c. The Federal Reserve Banks buy $2 billion of securities from commercial banks. Show the new balance-sheet numbers in column 3 of each table.

d. Now review each of the previous three transactions, asking yourself these three questions: (1) What change, if any, took place in the money supply as a direct and immediate result of each transaction? (2) What increase or decrease in the commercial banks' reserves took place in each transaction? (3) Assuming a reserve ratio of 20 percent, what change in the money-creating potential of the commercial banking system occurred as a result of each transaction?

4. Refer to Table 34.2 and assume that the Fed's reserve ratio is 10 percent and the economy is in a severe recession. Also suppose that the commercial banks are hoarding all excess reserves (not lending them out) because of their fear of loan defaults. Finally, suppose that the Fed is highly concerned that the banks will suddenly lend out these excess reserves and possibly contribute to inflation once the economy begins to recover and confidence is restored. By how many percentage points would the Fed need to increase the reserve ratio to eliminate one-third of the excess reserves? What would be the size of the monetary multiplier before and after the change in the reserve ratio? By how much would the lending potential of the banks decline as a result of the increase in the reserve ratio? **LO34.3**

5. Suppose that the demand for federal funds curve is such that the quantity of funds demanded changes by $120 billion for each 1 percent change in the federal funds interest rate. Also, assume that the current federal funds rate is at the 3 percent rate that is targeted by the Fed. Now suppose that the Fed retargets the rate to 3.5 percent. Assuming no change in demand, will the Fed need to increase or decrease the supply of federal funds? By how much will the quantity of federal funds have to change for the equilibrium to occur at the new target rate? **LO34.4**

6. Suppose that inflation is 2 percent, the federal funds rate is 4 percent, and real GDP falls 2 percent below potential GDP. According to the Taylor rule, in what direction and by how much should the Fed change the real federal funds rate? **LO34.4**

7. Refer to the table for Moola at the bottom of this page to answer the following questions. What is the equilibrium interest rate in Moola? What is the level of investment at the equilibrium interest rate? Is there either a recessionary output gap (negative GDP gap) or an inflationary output gap (positive GDP gap) at the equilibrium interest rate and, if either, what is the amount? Given money demand, by how much would the Moola central bank need to change the money supply to close the output gap? What is the expenditure multiplier in Moola? **LO34.5**

Money Supply	Money Demand	Interest Rate	Investment at Interest (Rate Shown)	Potential Real GDP	Actual Real GDP at Interest (Rate Shown)
$500	$800	2%	$50	$350	$390
500	700	3	40	350	370
500	600	4	30	350	350
500	500	5	20	350	330
500	400	6	10	350	310

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **35**

Financial Economics

Learning Objectives

LO35.1 Define *financial economics* and distinguish between economic investment and financial investment.

LO35.2 Explain the time value of money and how compound interest can be used to calculate the present value of any future amount of money.

LO35.3 Identify and distinguish between the most common financial investments: stocks, bonds, and mutual funds.

LO35.4 Relate how percentage rates of return provide a common framework for comparing assets and explain why asset prices and rates of return are inversely related.

LO35.5 Define and utilize the concept of arbitrage.

LO35.6 Describe how the word *risk* is used in financial economics and explain the difference between diversifiable and nondiversifiable risk.

LO35.7 Convey why investment decisions are determined primarily by investment returns and nondiversifiable risk and how investment returns compensate for being patient and for bearing nondiversifiable risk.

LO35.8 Explain how the Security Market Line illustrates the compensation that investors receive for time preference and nondiversifiable risk and why arbitrage will tend to move all assets onto the Security Market Line.

Financial economics studies investor preferences and how they affect the trading and pricing of financial assets like stocks, bonds, and real estate. The two most important investor preferences are a desire for high rates of return and a dislike of risk and uncertainty. This chapter will explain how these preferences interact to produce a strong positive relationship between risk and return: the riskier an investment, the higher its rate of return. This positive relationship compensates investors for bearing risk. And it is enforced by a powerful set of buying and selling pressures known as arbitrage, which ensures consistency across investments so that assets with identical levels of risk generate identical rates of return. As we will demonstrate, this consistency makes it extremely difficult for anyone to "beat the market" by finding a set of investments that can generate high rates of return at low levels of risk. Instead, investors are stuck with a trade-off: If they want higher rates of return, they must accept higher levels of risk. On average, higher risk results in higher returns. But it can also result in large losses, as it did for many investors who held risky assets during the financial crisis of 2007–2008.

Financial Investment

LO35.1 Define *financial economics* and distinguish between economic investment and financial investment.

Financial economics focuses its attention on the investments that individuals and firms make in the wide variety of assets available to them in our modern economy. But before proceeding, it is important for you to recall the difference between economic investment and financial investment.

Economic investment refers either to paying for *new* additions to the capital stock or *new* replacements for capital stock that has worn out. Thus, *new* factories, houses, retail stores, construction equipment, and wireless networks are all good examples of economic investments. And so are purchases of office computers to replace computers that have become obsolete as well as purchases of new commercial airplanes to replace planes that have served out their useful lives.

In contrast, financial investment is a far broader, much more inclusive concept. It includes economic investment and a whole lot more. **Financial investment** refers to either buying an asset or building an asset in the expectation of financial gain. It does not distinguish between *new* assets and *old* assets. Purchasing an old house or an old factory is just as much a financial investment as purchasing a new house or a new factory. For financial investment, it does not matter if the purchase of an asset adds to the capital stock, replaces the capital stock, or does neither. Investing in old comic books is just as much a financial investment as building a new refinery. Finally, unlike economic investment, financial investment can involve either *financial assets* (such as stocks, bonds, and futures contracts) or *real assets* (such as land, factories, and retail stores).

When bankers, entrepreneurs, corporate executives, retirement planners, and ordinary people use the word *investment*, they almost always mean financial investment. In fact, the ordinary meaning of the word investment is financial investment. So for this chapter, we will use the word investment in its ordinary sense of "financial investment" rather than in the far narrower sense of "economic investment," which is used throughout the rest of this book.

Present Value

LO35.2 Explain the time value of money and how compound interest can be used to calculate the present value of any future amount of money.

Money has "time value" because current dollars can be converted into a larger amount of future dollars through compound interest. The *time-value of money* is the idea that a specific amount of money is more valuable to a person the sooner it is received, and a person will need to be compensated for waiting to obtain it later. The time-value of money can also be thought of as the opportunity cost of receiving a sum of money later rather than earlier.

The time-value of money underlies one of the most fundamental ideas in financial economics: **present value,** which is the present-day value, or worth, of returns or costs that are expected to arrive in the future. The ability to calculate present values is especially useful when investors wish to determine the proper current price to pay for an asset. In fact, the proper current price for any risk-free investment *is* the present value of its expected future returns. And while some adjustments have to be made when determining the proper price of a risky investment, the process is entirely based on the logic of present value. So we begin our study of finance by explaining present value and how it can be used to price risk-free assets. Once that is accomplished, we will turn our attention to risk and how the financial markets determine the prices of risky assets by taking into account investor preferences regarding the trade-off between potential return and potential risk.

Compound Interest

The best way to understand present value is by first understanding compound interest. **Compound interest** describes how quickly an investment increases in value when interest is paid, or compounded, not only on the original amount invested but also on all interest payments that have been previously made.

As an example of compound interest in action, consider Table 35.1, which shows the amount of money that \$100 invested today becomes if it increases, or compounds, at an 8 percent annual interest rate, i, for various numbers of years. To simplify, let's express the 8 percent annual interest rate as a decimal so that it becomes $i = 0.08$. The key to understanding compound interest is to realize that 1 year's worth of growth at interest rate i will always result in $(1 + i)$ times as much money at the end of a year as there was at the beginning of the year. Consequently, if the first year begins with \$100 and if $i = 0.08$, then $(1 + 0.08)$ or 1.08 times as much money—\$108—will be available at the end of the year. We show the computation for the first year in column 2 of Table 35.1 and display the \$108 outcome in column 3. The same logic would also apply with other initial amounts. If a year begins with \$500, there will be 1.08 times more money after 1 year, or \$540. Algebraically, let X_0 denote the amount of money at the start of the first year and X_1 the amount after one year's worth of growth. Then we see that any given number of dollars X_0 at the start of the first year grows into $X_1 = (1 + i)X_0$ dollars after one year's worth of growth.

Next, consider what happens if the initial investment of \$100 that grew into \$108 after 1 year continues to grow at 8 percent interest for a second year. The \$108 available at the beginning of the second year will grow into an amount of money that is 1.08 times larger by the end of the second year. That amount, as shown in Table 35.1, is \$116.64. Notice that the computation in the table is made by multiplying the initial \$100 by $(1.08)^2$. That is because the original \$100 is compounded by 1.08 into \$108 and then the \$108 is again compounded by 1.08. More generally, since the second year begins with $(1 + i)X_0$ dollars, it will grow to $(1 + i)(1 + i)X_0 = (1 + i)^2X_0$ dollars by the end of the second year.

TABLE 35.1 Compounding: \$100 at 8 Percent Interest

(1) Years of Compounding	(2) Compounding Computation	(3) Value at Year's End
1	\$100 (1.08)	\$108.00
2	100 $(1.08)^2$	116.64
3	100 $(1.08)^3$	125.97
4	100 $(1.08)^4$	136.05
5	100 $(1.08)^5$	146.93
17	100 $(1.08)^{17}$	370.00

Similar reasoning shows that the amount of money at the end of 3 years has to be $(1 + i)^3X_0$ since the amount of money at the beginning of the third year, $(1 + i)^2X_0$, gets multiplied by $(1 + i)$ to convert it into the amount of money at the end of the third year. In terms of Table 35.1, that amount is \$125.97, which is $(1.08)^3$\$100.

As you can see, we now have a fixed pattern. The \$100 that is invested at the beginning of the first year becomes $(1 + i)$\$100 after 1 year, $(1 + i)^2$\$100 after 2 years, $(1 + i)^3$\$100 after 3 years, and so on. It therefore is clear that the amount of money after t years will be $(1 + i)^t$\$100. This pattern always holds true, regardless of the size of the initial investment. Thus, investors know that if X_0 dollars is invested today and earns compound interest at the rate i, it will grow into exactly $(1 + i)^tX_0$ dollars after t years. Economists express that fact with the following formula:

$$X_t = (1 + i)^t X_0 \qquad (1)$$

Equation 1 captures the idea that if investors have the opportunity to invest X_0 dollars today at interest rate i, then they have the ability to transform X_0 dollars today into $(1 + i)^tX_0$ dollars in t years.

But notice that the logic of the equality also works in reverse, so that it can also be thought of as showing that $(1 + i)^tX_0$ dollars in t years can be transformed into X_0 dollars today. That may seem very odd, but it is exactly what happens when people take out loans. For instance, consider a situation where an investor named Roberto takes out a loan for \$100 dollars today, a loan that will accumulate interest at 8 percent per year for 5 years. Under such an arrangement, the amount Roberto owes will grow with compound interest into $(1.08)^5$\$100 = \$146.93 dollars in 5 years. This means that Roberto can convert \$146.93 dollars in 5 years (the amount required to pay off the loan) into \$100 dollars today (the amount he borrows).

Consequently, the compound interest formula given in equation 1 defines not only the rate at which present amounts of money can be converted to future amounts of money but also the rate at which future amounts of money can be converted into present amounts of money.

The Present Value Model

The present value model simply rearranges equation 1 to make it easier to transform future amounts of money into present amounts of money. To derive the formula used to calculate the present value of a future amount of money, we divide both sides of equation 1 by $(1 + i)^t$ to obtain

$$\frac{X_t}{(1 + i)^t} = X_0 \qquad (2)$$

The logic of equation 2 is identical to that of equation 1. Both allow investors to convert present amounts of money into future amounts of money and vice versa. However, equation 2 makes it much more intuitive to convert a given number of dollars in the future into their present-day equivalent. In fact, it says that X_t dollars in t years converts into exactly $X_t/(1 + i)^t$ dollars today. This may not seem important, but it is actually very powerful because it allows investors to easily calculate how much they should pay for any given asset.

WORKED PROBLEMS

W35.1

Present value

To see why this is true, understand that an asset's owner obtains the right to receive one or more future payments. If an investor is considering buying an asset, her problem is to try to determine how much she should pay today to buy the asset and receive those future payments. Equation 2 makes this task very easy. If she knows how large a future payment will be (X_t dollars), when it will arrive (in t years), and what the interest rate (i) is, then she can apply equation 2 to determine the payment's present value: its value in present-day dollars. If she does this for each of the future payments that the asset in question is expected to make, she will be able to calculate the overall present value of all the asset's future payments by simply summing together the present values of each of the individual payments. This will allow her to determine the price she should pay for the asset. In particular, *the asset's price should exactly equal the sum of the present values of all of the asset's future payments*.

As a simple example, suppose that Cecilia has the chance to buy an asset that is guaranteed to return a single payment of exactly \$370.00 in 17 years. Again let's assume the interest rate is 8 percent per year. Then the present value of that future payment can be determined using equation 2 to equal precisely $\$370.00/(1 + 0.08)^{17} = \$370.00/(1.08)^{17} = \$100$ today. This is confirmed in the row for year 17 in Table 35.1.

To see why Cecilia should be willing to pay a price that is *exactly* equal to the \$100 present value of the asset's single future payment of \$370.00 in 17 years, consider the following thought experiment. What would happen if she were to invest \$100 today in an alternative investment that is guaranteed to compound her money for 17 years at 8 percent per year? How large would her investment in this alternative become? Equation 1 and Table 35.1 tell us that the answer is exactly \$370.00 in 17 years.

This is very important because it shows that Cecelia and other investors have two different possible ways of purchasing the right to receive \$370.00 in 17 years. They can either:

- Purchase the asset in question for \$100.
- Invest \$100 in the alternative asset that pays 8 percent per year.

Because either investment will deliver the same future benefit, both investments are in fact identical. Consequently, they should have identical prices—meaning that each will cost precisely \$100 today.

A good way to see why this must be the case is by considering how the presence of the alternative investment affects the behavior of both the potential buyers and the potential sellers of the asset in question. First, notice that Cecelia and other potential buyers would never pay more than \$100 for the asset in question because they know that they could get the same future return of \$370.00 in 17 years by investing \$100 in the alternative investment. At the same time, people selling the asset in question would not sell it to Cecelia or other potential buyers for anything less than \$100 since they know that the only other way for Cecelia and other potential buyers to get a future return of \$370.00 in 17 years is by paying \$100 for the alternative investment. Since Cecelia and the other potential buyers will not pay more than \$100 for the asset in question and its sellers will not accept less than \$100 for the asset in question, the result will be that the asset in question and the alternative investment will have the exact same price of \$100 today.

QUICK REVIEW 35.1

- Financial investment refers to buying an asset with the hope of financial gain.
- The time-value of money is the idea that a specific amount of money is more valuable to a person the sooner it is received because of the potential for compound interest.
- Compound interest is the payment of interest not only on the original amount invested but also on any interest payments previously made; X_0 dollars today growing at interest rate i will become $(1 + i)^t X_0$ dollars in t years.
- The present value formula facilitates transforming future amounts of money into present-day amounts of money; X_t dollars in t years converts into exactly $X_t/(1 + i)^t$ dollars today.
- An investment's proper current price is equal to the sum of the present values of all the future payments that it is expected to make.

Applications

Present value is not only an important idea for understanding investment, but it has many everyday applications. Let's examine two of them.

Take the Money and Run? The winners of state lotteries are typically paid their winnings in equal installments spread out over 20 years. For instance, suppose that Zoe gets lucky one week and wins a $100 million jackpot. She will not be paid $100 million all at once. Rather, she will receive $5 million per year for 20 years, for a total of $100 million.

Zoe may object to this installment payment system for a variety of reasons. For one thing, she may be very old, so that she is not likely to live long enough to collect all of the payments. Alternatively, she might prefer to receive her winnings immediately so that she could make large immediate donations to her favorite charities or large immediate investments in a business project that she would like to get started. And, of course, she may just be impatient and want to buy a lot of really expensive consumption goods sooner rather than later.

Fortunately for Zoe, if she does have a desire to receive her winnings sooner rather than later, several private financial companies are ready and willing to help her. They do this by arranging swaps. Lottery winners sell the right to receive their installment payments in exchange for a single lump sum that they get immediately. The people who hand over the lump sum receive the right to collect the installment payments.

Present value is crucial to arranging these swaps since it is used to determine the value of the lump sum that lottery winners like Zoe will receive in exchange for giving up their installment payments. The lump sum in any case is simply equal to the sum of the present values of each of the future payments. Assuming an interest rate of 5 percent per year, the sum of the present values of each of Zoe's 20 installment payments of $5 million is $62,311,051.71. So, depending on her preferences, Zoe can either receive that amount immediately or $100 million spread out over 20 years.

Salary Caps and Deferred Compensation Another example of present value comes directly from the sporting news. Many professional sports leagues worry that richer teams, if not held in check, would outbid poorer teams for the best players. The result would be a situation in which only the richer teams have any real chance of doing well and winning championships.

To prevent this from happening, many leagues have instituted salary caps. These are upper limits on the total amount of money that each team can spend on salaries during a given season. For instance, one popular basketball league has a salary cap of about $58 million per season, so that the combined value of the salaries that each team pays its players can be no more than $58 million.

Typically, however, the salary contracts that are negotiated between individual players and their teams are for multiple seasons. This means that during negotiations, players are often asked to help their team stay under the current season's salary cap by agreeing to receive more compensation in later years. For instance, suppose that a team's current payroll is $53 million but that it would like to sign a superstar nicknamed HiTop to a two-year contract. HiTop, however, is used to earning $10 million per year. This is a major problem for the team because the $58 million salary cap means that the most that the team can pay HiTop for the current season is $5 million.

A common solution is for HiTop to agree to receive only $5 million the first season in order to help the team stay under the salary cap. In exchange for this concession, the team agrees to pay HiTop more than the $10 million he would normally demand for the second season. The present value formula is used to figure out how large his second-season salary should be. In particular, the player can use the present value formula to figure out that if the interest rate is 8 percent per year, he should be paid a total of $15,400,000 during his second season, since this amount will equal the $10 million he wants for the second season plus $5.4 million to make up for the $5 million reduction in his salary during the first season. That is, the present value of the $5.4 million that he will receive during the second season precisely equals the $5 million that he agrees to give up during the first season.

Some Popular Investments

LO35.3 Identify and distinguish between the most common financial investments: stocks, bonds, and mutual funds.

The number and types of financial "instruments" in which one can invest are very numerous, amazingly creative, and highly varied. Most are much more complicated than the investments we used to explain compounding and present value. But, fortunately, all investments share three features:

- They require that investors pay some price—determined in the market—to acquire them.
- They give their owners the chance to receive future payments.
- The future payments are typically risky.

These features allow us to treat all assets in a unified way. Three of the more popular investments are stocks, bonds, and mutual funds. In 2010, the median value of stock holdings for U.S. families that held stocks was $17,800; the median value for bonds, $83,800; and the median value for "pooled funds" (mainly mutual funds) was $58,700.[1]

Stocks

Recall that **stocks** are ownership shares in a corporation. If an investor owns 1 percent of a corporation's shares, she gets 1 percent of the votes at the shareholder meetings that select the company's managers and she is also entitled to 1 percent of any future profit distributions. There is no guarantee, however, that a company will be profitable.

Firms often lose money and sometimes even go **bankrupt,** meaning that they are unable to make timely payments on their debts. In the event of a bankruptcy, control of a corporation's assets is given to a bankruptcy judge, whose job is to enforce the legal rights of the people who lent the company money by doing what he can to see that they are repaid. Typically, this involves selling off the corporation's assets (factories, real estate, patents, etc.) to raise the money necessary to pay off the company's debts. The money raised by selling the assets may be greater than or less than what is needed to fully pay off the firm's debts. If it is more than what is necessary, any remaining money is divided equally among shareholders. If it is less than what is necessary, then the lenders do not get repaid in full and have to suffer a loss.

A key point, however, is that the maximum amount of money that shareholders can lose is what they pay for their shares. If the company goes bankrupt owing more than the value of the firm's assets, shareholders do not have to make up the difference. This **limited liability rule** limits the risks involved in investing in corporations and encourages investors to invest in stocks by capping their potential losses at the amount that they paid for their shares.

When firms are profitable, however, investors can look forward to gaining financially in either or both of two possible ways. The first is through **capital gains,** meaning that they sell their shares in the corporation for more money than they paid for them. The second is by receiving **dividends,** which are equal shares of the corporation's profits. As we will soon explain, a corporation's current share price is determined by the size of the capital gains and dividends that investors expect the corporation to generate in the future.

[1]Federal Reserve, "Changes in U.S. Family Finances from 2007–2010; Evidence from the Survey of Consumer Finances," p. 26.

Bonds

Bonds are debt contracts that are issued most frequently by governments and corporations. They typically work as follows: An initial investor lends the government or the corporation a certain amount of money, say $1,000, for a certain period of time, say 10 years. In exchange, the government or corporation promises to make a series of semiannual payments in addition to returning the $1,000 at the end of the 10 years. The semiannual payments constitute interest on the loan. For instance, the bond agreement may specify that the borrower will pay $30 every six months. This means that the bond will pay $60 per year in payments, which is equivalent to a 6 percent rate of interest on the initial $1,000 loan.

The initial investor is free, however, to sell the bond at any time to other investors, who then gain the right to receive any of the remaining semiannual payments as well as the final $1,000 payment when the bond expires after 10 years. As we will soon demonstrate, the price at which the bond will sell if it is indeed sold to another investor will depend on the current rates of return available on other investments offering a similar stream of future payments and facing a similar level of risk.

The primary risk a bondholder faces is the possibility that the corporation or government that issues his bond will **default** on, or fail to make, the bond's promised payments. This risk is much greater for corporations, but it also faces local and state governments in situations where they cannot raise enough tax revenue to make their bond payments or where defaulting on bond payments is politically easier than reducing spending on other items in the government's budget to raise the money needed to keep making bond payments. The U.S. federal government, however, has never defaulted on its bond payments and is very unlikely to ever default because it has access to huge amounts of current and potential tax revenue and can sell U.S. securities to the Fed as a way to obtain money.

A key difference between bonds and stocks is that bonds are much more predictable. Unless a bond goes into default, its owner knows both how big its future payments will be and exactly when they will arrive. By contrast, stock prices and dividends are highly volatile because they depend on profits, which vary greatly depending on the overall business cycle and on factors specific to individual firms and industries—things such as changing consumer preferences, variations in the costs of inputs, and changes in the tax code. As we will demonstrate later, the fact that bonds are typically more predictable (and thus less risky) than stocks explains why they generate lower average rates of return than stocks. Indeed, this difference in rates of return has been very large historically. From 1926 to 2012, stocks on average returned about 11 percent per year worldwide, while bonds on average returned only roughly 6 percent per year worldwide.

Mutual Funds

A **mutual fund** is a company that maintains a professionally managed **portfolio,** or collection, of either stocks or bonds. The portfolio is purchased by pooling the money of many investors. Since these investors provide the money to purchase the portfolio, they own it and any gains or losses generated by the portfolio flow directly to them. Table 35.2 lists the 10 largest U.S. mutual funds based on their assets.

Most of the more than 8,000 mutual funds currently operating in the United States choose to maintain portfolios that invest in specific categories of bonds or stocks. For instance, some fill their portfolios exclusively with the stocks of small tech companies, while others buy only bonds issued by certain state or local governments. In addition, there are **index funds,** whose portfolios are selected to exactly match a stock or bond index. Indexes follow the performance of a particular group of stocks or bonds to gauge how well a particular category of investments is doing. For instance, the Standard & Poor's 500 Index contains the 500 largest stocks trading in the United States to capture how the stocks of large corporations vary over time, while the Lehman 10-Year Corporate Bond Index follows a representative collection of 10-year corporate bonds to see how well corporate bonds do over time.

An important distinction must be drawn between actively managed and passively managed mutual funds. **Actively managed funds** have portfolio managers who constantly buy and sell assets in an attempt to generate high returns. By contrast, index funds are **passively managed funds** because the assets in their portfolios are chosen to exactly match whatever stocks or bonds are contained in their respective underlying indexes.

Later in the chapter, we will discuss the relative merits of actively managed funds and index funds, but for now we merely point out that both types are very popular and that, overall, U.S. households and nonprofit organizations held $5.3 trillion in mutual funds at the end of 2012. By way of comparison, U.S. GDP in 2012 was $15.7 trillion and the estimated value of all the financial assets held by households and nonprofit organizations in 2012 (including everything from individual stocks and bonds to checking account deposits) was about $54 trillion.

TABLE 35.2 **The 10 Largest Mutual Funds, March 2013**

Fund Name*	Assets under Management, Billions
PIMCO: Total Return Institutional	$179.9
SPDR S&P 500 ETF	129.8
Vanguard Total Stock Index Investor	90.1
Vanguard Institutional Index	75.2
Vanguard Total Stock Index Admiral Shares	68.0
Vanguard 500 Index Admiral Shares	66.5
Fidelity Contrafund	63.4
SPDR Gold	62.7
American Funds Income A	61.5
American Funds Capital Income Builder A	61.1

*The letter *A* indicates funds that have sales commissions and are generally purchased by individuals through their financial advisors.

Source: Lipper Performance Report, March 31, 2013.

QUICK REVIEW 35.2

- Three popular forms of financial investments are stocks, bonds, and mutual funds.
- Stocks are ownership shares in corporations and bestow upon their owners a proportional share of any future profit.
- Bonds are debt contracts that promise to pay a fixed series of payments in the future.
- Mutual funds are pools of investor money used to buy a portfolio of stocks or bonds.

Calculating Investment Returns

LO35.4 Relate how percentage rates of return provide a common framework for comparing assets and explain why asset prices and rates of return are inversely related.

Investors buy assets to obtain one or more future payments. The simplest case is purchasing an asset for resale. For instance, an investor may buy a house for $300,000 with the hope of selling it for $360,000 in one year. On the other hand, he could also rent out the house for $3,000 per month and thereby receive a stream of future payments. And he, of course, could do a little of both, paying $300,000 for the house now to rent it out for five years and then sell it. In that case, he is expecting a stream of smaller payments followed by a large one.

Percentage Rates of Return

Economists have developed a common framework for evaluating the gains or losses of assets that only make one future payment as well as those that make many future payments.

They state the gain or loss as a **percentage rate of return,** by which they mean the percentage gain or loss (relative to the buying price) over a given period of time, typically a year. For instance, if Noelle buys a rare comic book today for $100 and sells it in 1 year for $125, she is said to make a 25 percent per year rate of return because she would divide the gain of $25 by the purchase price of $100. By contrast, if she were only able to sell it for $92, then she would be said to have made a loss of 8 percent per year since she would divide the $8 loss by the purchase price of $100.

A similar calculation is made for assets that deliver a series of payments. For instance, an investor who buys a house for $300,000 and expects to rent it out for $3,000 per month would be expecting to make a 12 percent per year rate of return because he would divide his $36,000 per year of rent by the $300,000 purchase price of the house.

The Inverse Relationship between Asset Prices and Rates of Return

A fundamental concept in financial economics is that, other things equal, *an investment's rate of return is inversely related to its price*. The higher the price paid for a fixed-return asset, the lower the rate of return on the investment. To see why, consider a bond that pays $24,000 of interest each year. If an investor pays $100,000 for the bond, he will earn a 24 percent per year rate of return because the $24,000 annual interest payment will be divided by the $100,000 purchase price of the bond.

But suppose that the purchase price of the bond rises to $200,000. In that case, the investor would earn only a 12 percent per year rate of return, since the $24,000 annual interest payment would be divided by the much larger purchase price of $200,000. Consequently, as the price of the bond goes up, the rate of return from buying it goes down.

The same relationship holds for other fixed-return investments, such as rental property. The higher the price of the property, given its monthly rent, the lower the rate of return. The underlying cause of this general relationship is the fact that the annual payments are fixed in value so that there is an upper limit to the financial rewards of owning the asset. As a result, the more an investor pays for the asset, the lower the asset's rate of return.

Arbitrage

LO35.5 Define and utilize the concept of arbitrage.

Arbitrage is the name that financial economists give to the buying and selling process that leads profit-seeking investors to equalize the average expected rates of return generated by identical or nearly identical assets. Arbitrage happens when investors try to take advantage and profit from situations where two identical or nearly identical assets have different rates of return. They do so by simultaneously selling the asset with the lower rate of return and buying the asset with the higher rate of return. For instance, consider what would happen in a case where two very similar T-shirt companies start with different rates of return despite the fact that they are equally profitable and have equally good future prospects. To make things concrete, suppose that a company called T4me starts out with a rate of return of 10 percent per year while TSTG (T-Shirts to Go) starts out with a rate of return of 15 percent per year.

Since both companies are basically identical and have equally good prospects, investors in T4me will want to shift over to TSTG, which offers higher rates of return for the same amount of risk. As they begin to shift over, however, the prices of the two companies will change—and with them, the rates of return on the two companies. In particular, since so many investors will be selling the shares of the lower-return company, T4me, the supply of its shares trading on the stock market will rise so that its share price will fall. But since asset prices and rates of return are inversely related, this will cause its rate of return to rise.

At the same time, however, the rate of return on the higher-return company, TSTG, will begin to fall. This has to be the case because, as investors switch from T4me to TSTG, the increased demand for TSTG's shares will drive up their price. And as the price of TSTG goes up, its rate of return must fall.

The interesting thing is that this arbitrage process will continue—with the rate of return on the higher-return company falling and the rate of return on the lower-return company rising—until both companies have the same rate of return. This convergence must happen because as long as the rates of return on the two companies are not identical, there will always be some investors who will want to sell the shares of the lower-return company so they can buy the shares of the higher-return company. As a result, arbitrage will continue until the rates of return are equal.

What is even more impressive, however, is that generally only a very short while is needed for prices to equalize. In fact, for highly traded assets like stocks and bonds, arbitrage will often force the rates of return on identical or nearly identical investments to converge within a matter of minutes or sometimes even within a matter of seconds. This is very helpful to small investors who do not have a large amount of time to study the thousands of potential investment opportunities available in the financial markets.

Thanks to arbitrage, they can invest with the confidence that assets with similar characteristics will have similar rates of return. As we discuss in the next section, this is especially important when it comes to risk—a characteristic that financial economists believe investors care about very deeply.

QUICK REVIEW 35.3

- Stating gains or losses as percentage rates of return provides a common framework for comparing different assets.
- Asset prices and percentage rates of return are inversely related.
- Arbitrage refers to the buying and selling that takes place to equalize the percentage rates of return on identical or nearly identical assets.

Risk

LO35.6 Describe how the word *risk* is used in financial economics and explain the difference between diversifiable and nondiversifiable risk.

Investors purchase assets to obtain one or more future payments. As used by financial economists, the word **risk** refers to the fact that investors never know with total certainty what those future payments will turn out to be.

The underlying problem is that the future is uncertain. Many factors affect an investment's future payments, and each of these may turn out better or worse than expected. As a simple example, consider buying a farm. Suppose that in an average year, the farm will generate a profit of $100,000. But if a freak hailstorm damages the crops, the profit will fall to only $60,000. On the other hand, if weather conditions turn out to be perfect, the profit will rise to $120,000. Since there is no way to tell in advance what will happen, investing in the farm is risky.

Also notice that when financial economists use the word *risk*, they do not use it in the casual way that refers only to potentially bad outcomes (as in, "there is a risk that this experimental medicine may kill you"). In financial economics, the word *risk* means only that an outcome—good or bad, major or minor, likely or unlikely—lacks total certainty. Some outcome will occur, but you cannot be sure what it will be. For instance, suppose that you are gifted a raffle ticket that will pay you $10, $100, or $1,000 when a drawing is made in one month. There are no bad outcomes in this particular case, only good ones. But because you do not know with certainty which outcome will occur, the situation is, by definition, risky. On the other hand, the word *risk* in financial economics certainly does not preclude negative outcomes. If you buy shares of common stock in some company, your stock may go up in value. But the company could also go bankrupt, in which case you would lose your entire investment.

Diversification

Investors have many options regarding their portfolios, or collections of investments. Among other things, they can choose to concentrate their wealth in just one or two investments or spread it out over a large number of investments. **Diversification** is the name given to the strategy of investing in a large number of investments to reduce the overall risk to the entire portfolio.

The underlying reason that diversification generally succeeds in reducing risk is best summarized by the old saying, "Don't put all your eggs in one basket." If an investor's portfolio consists of only one investment, say one stock, then if anything awful happens to that stock, the investor's entire portfolio will suffer greatly. By contrast, if the investor spreads his wealth over many stocks, then a bad outcome for any one particular stock will cause only a small amount of damage to the overall portfolio. In addition, it will typically be the case that if something bad is happening to one part of the portfolio, something good will be happening to another part of the portfolio and the two effects will tend to offset each other. Thus, the risk to the overall portfolio is reduced by diversification.

It must be stressed, however, that while diversification can reduce a portfolio's risks, it cannot eliminate them entirely. The problem is that even if an investor has placed each of his eggs into a different basket, all of the eggs may still end up broken if all of the different baskets somehow happen to get dropped simultaneously. That is, even if an investor has created a well-diversified portfolio, all of the investments still have a chance to do badly simultaneously. As an example, consider the early portion of the severe recession of 2007–2009: With economic activity declining and consumer spending falling, nearly all companies faced reduced sales and lowered profits, a fact that caused their stock prices to decline simultaneously. Consequently, even if investors had diversified their portfolios across numerous stocks, their overall wealth portfolios would have still declined because nearly all of their many investments simultaneously performed poorly.

ORIGIN OF THE IDEA

O35.1

Portfolio diversification

Financial economists build on the intuition behind the benefits and limits to diversification to divide an individual investment's overall risk into two components, diversifiable risk and nondiversifiable risk. **Diversifiable risk** (or "idiosyncratic risk") is the risk that is specific to a given investment and that can be eliminated by diversification. For instance, a soda pop maker faces the risk that the demand for its product may suddenly decline because people will want to drink mineral water instead of soda pop. But this risk does not matter if an investor has a diversified portfolio that contains stock in the soda pop maker as well as stock in a mineral water maker. This is true because when the stock price of the soda pop maker falls due to the change in consumer preferences, the stock price of the mineral water maker will go up—so that, as far as the overall portfolio is concerned, the two effects will offset each other.

By contrast, **nondiversifiable risk** (or "systemic risk") pushes all investments in the same direction at the same time so that there is no possibility of using good effects to offset bad effects. The best example of a nondiversifiable risk is the business cycle. If the economy does well, then corporate profits rise and nearly every stock does well. But if the economy does badly, then corporate profits fall and nearly every stock does badly. As a result, even if one were to build a well-diversified portfolio, it would still be affected by the business cycle because nearly every asset contained in the portfolio would move in the same direction at the same time whenever the economy improved or worsened.

That being said, creating a diversified portfolio is still an investor's best strategy because doing so at least eliminates diversifiable risk. Indeed, it should be emphasized that for investors who have created diversified portfolios, all diversifiable risks will be eliminated, so that the only remaining source of risk will be nondiversifiable risk.

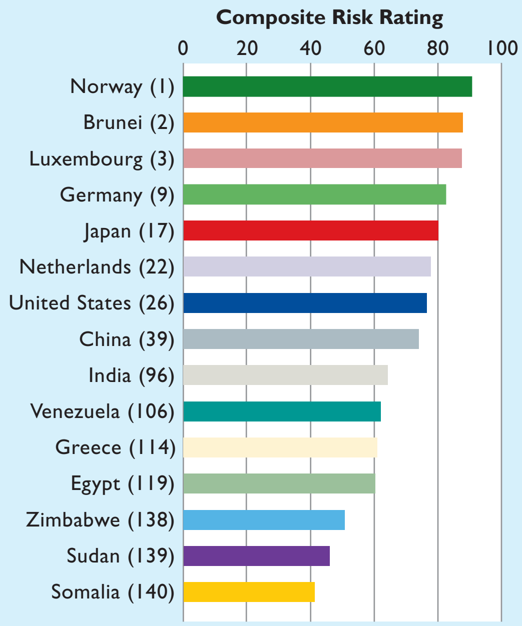
A significant implication of this fact is that when investors consider whether to add any particular investment to a portfolio that is already diversified, they can ignore the investment's diversifiable risk. They can ignore it because, as part of a diversified portfolio, the investment's diversifiable risk will be "diversified away." Indeed, the only risk left will be the amount of nondiversifiable risk that the investment carries with it. This is crucial because it means that investors can base their decisions about whether to add a potential new investment to their portfolios on a comparison between the potential investment's level of nondiversifiable risk and its potential returns. If they find this trade-off attractive, they will add the investment, whereas if it seems unattractive, they will not.

The next section shows how investors can measure each asset's level of nondiversifiable risk as well as its potential returns to facilitate such comparisons. Global Perspective 35.1 shows how investment risk varies significantly across countries due to underlying differences in political, economic, and financial risks.

GLOBAL PERSPECTIVE 35.1

Investment Risks Vary across Different Countries

The International Country Risk Guide is a monthly publication that attempts to distill the political, economic, and financial risks facing 140 countries into a single "composite risk rating" number for each country, with higher numbers indicating less risk and more safety. The table below presents the July 2012 ranks and rating numbers for 15 countries including the three least risky (ranked 1 through 3) and the three most risky (ranked 138 through 140.) Risk ratings numbers above 80 are considered *very low risk;* 70–80 are considered *low risk;* 60–70 *moderate risk;* 50–60 *high risk;* and below 50 *very high risk*.



Source: *International Country Risk Guide,* a publication of The PRS Group, Inc., **www.prsgroup.com**, 2012. Used with permission.

QUICK REVIEW 35.4

- An asset is risky if its future payments are uncertain.
- Diversification is an investment strategy in which an investor invests in a large number of different investments in order to reduce the overall risk to his or her entire portfolio.
- Risks that can be canceled out by diversification are called diversifiable risks; risks that cannot be canceled out by diversification are called nondiversifiable risks.

Comparing Risky Investments

LO35.7 Convey why investment decisions are determined primarily by investment returns and nondiversifiable risk and how investment returns compensate for being patient and for bearing nondiversifiable risk.

Economists believe that the two most important factors affecting investment decisions are returns and risk—specifically nondiversifiable risk. But for investors to properly compare different investments on the basis of returns and risk, they need ways to measure returns and risk. The two standard measures are, respectively, the average expected rate of return and the beta statistic.

Average Expected Rate of Return

Each investment's **average expected rate of return** is the probability-weighted average of the investment's possible future rates of return. The term **probability-weighted average** simply means that each of the possible future rates of return is multiplied by its probability expressed as a decimal (so that a 50 percent probability is 0.5 and a 23 percent probability is 0.23) before being added together to obtain the average. For instance, if an investment has a 75 percent probability of generating 11 percent per year and a 25 percent probability of generating 15 percent per year, then its average expected rate of return will be 12 percent = (0.75 × 11 percent) + (0.25 × 15 percent). By weighting each possible outcome by its probability, this process ensures that the resulting average gives more weight to those outcomes that are more likely to happen (unlike the normal averaging process that would treat every outcome the same).

Once investors have calculated the average expected rates of return for all the assets they are interested in, there will naturally be some impulse to simply invest in those assets having the highest average expected rates of return. But while this might satisfy investor cravings for higher rates of return, it would not take proper account of the fact that investors dislike risk and uncertainty. To quantify their dislike, investors require a statistic that can measure each investment's risk level.

Beta

One popular statistic that measures risk is called beta. **Beta** is a *relative* measure of nondiversifiable risk. It measures how the nondiversifiable risk of a given asset or portfolio of assets compares with that of the **market portfolio,** which is the name given to a portfolio that contains every asset available in the financial markets. The market portfolio is a useful standard of comparison because it is as diversified as possible. In fact, since it contains every possible asset, every possible diversifiable risk will be diversified away—meaning that it will be exposed *only* to nondiversifiable risk. Consequently, it can serve as a useful benchmark against which to measure the levels of nondiversifiable risk to which individual assets are exposed.

Such comparisons are very simple because the beta statistic is standardized such that the market portfolio's level of nondiversifiable risk is set equal to 1.0. Consequently, an asset with beta = 0.5 has a level of nondiversifiable risk that is one-half of that possessed by the market portfolio, while an asset with beta = 2.0 has twice as much nondiversifiable risk as the market portfolio. In addition, the beta numbers of various assets also can be used to compare them with each other. For instance, an asset with beta = 2.0 has four times as much exposure to nondiversifiable risk as does an asset with beta = 0.5.

Another useful feature of beta is that it can be calculated not only for individual assets but also for portfolios. Indeed, it can be calculated for portfolios no matter how many or how few assets they contain and no matter what those assets happen to be. This fact is very convenient for mutual fund investors because it means that they can use beta to quickly see how the nondiversifiable risk of any given fund's portfolio compares with that of other potential investments that they may be considering.

The beta statistic is used along with average expected rates of return to give investors standard measures of risk and return that can be used to sensibly compare different investment opportunities. As we will discuss in the next section, this leads to one of the most fundamental relationships in financial economics: riskier assets have higher rates of return.

Relationship of Risk and Average Expected Rates of Return

The fact that investors dislike risk has a profound effect on asset prices and average expected rates of return. In particular, their dislike of risk and uncertainty causes investors to pay higher prices for less-risky assets and lower prices for more-risky assets. But since asset prices and average expected rates of return are inversely related, this implies that less risky assets will have lower average expected rates of return than more risky assets.

Stated a bit more clearly: *Risk levels and average expected rates of return are positively related*. The more risky an investment is, the higher its average expected rate of return will be. A great way to understand this relationship is to think of higher average expected rates of return as being a form of compensation. Since investors dislike risk, they demand higher levels of compensation the more risky an asset is. The higher levels of compensation come in the form of higher average expected rates of return.

Be sure to note that this phenomenon affects all assets. Regardless of whether the assets are stocks or bonds or real estate or anything else, assets with higher levels of risk always end up with higher average expected rates of return to compensate investors for the higher levels of risk involved. No matter what the investment opportunity is, investors examine its possible future payments, determine how risky they are, and then select a price that reflects those risks. Since less-risky investments get higher prices, they end up with lower rates of return, whereas more-risky investments end up with lower prices and, consequently, higher rates of return.

The Risk-Free Rate of Return

We have just shown that there is a positive relationship between risk and returns, with higher returns serving to compensate investors for higher levels of risk. One investment, however, is considered to be risk-free for all intents and purposes. That investment is short-term U.S. government bonds.

These bonds are short-term loans to the U.S. government, with the duration of the loans ranging from 4 weeks to 26 weeks. They are considered to be essentially risk-free because there is almost no chance that the U.S. government will not be able to repay these loans on time and in full. To pay its debt, it can shift funds within its enormous budget, raise taxes, or borrow newly created money from the Fed. Although it is true that the U.S. government may eventually be destroyed or disabled to such an extent that it will not be able to repay some of its loans, the chances of such a calamity happening within 4 or even 26 weeks are essentially zero. Consequently, because it is a near certainty that the bonds will be repaid in full and on time, they are considered by investors to be risk-free.

Since higher levels of risk lead to higher rates of return, a person might be tempted to assume—incorrectly—that since government bonds are risk-free, they should earn a zero percent rate of return. The problem with this line of thinking is that it mistakenly assumes that risk is the *only* thing that rates of return compensate for. The truth is that rates of return compensate not only for risk but also for something that economists call time preference.

Time preference refers to the fact that because people tend to be impatient, they typically prefer to consume things in the present rather than in the future. Stated more concretely, most people, if given the choice between a serving of their favorite dessert immediately or a serving of their favorite dessert in five years, will choose to consume their favorite dessert immediately.

This time preference for consuming sooner rather than later affects the financial markets because people want to be compensated for delayed consumption. In particular, if Dave asks Oprah to lend him \$1 million for one year, he is implicitly asking Oprah to delay consumption for a year because if she lends Dave the \$1 million, she will not be able to spend that money herself for at least a year. If Oprah is like most people and has a preference for spending her \$1 million sooner rather than later, the only way Dave will be able to convince Oprah to let him borrow \$1 million is to offer her some form of compensation. The compensation comes in the form of an interest payment that will allow Oprah to consume more in the future than she can now. For instance, Dave can offer to pay Oprah \$1.1 million in one year in exchange for \$1 million today. That is, Oprah will get back the \$1 million she lends to Dave today as well as an extra \$100,000 to compensate her for being patient.

Notice the very important fact that this type of interest payment has nothing to do with risk. It is purely compensation for being patient and must be paid even if there is no risk involved and 100 percent certainty that Dave will fulfill his promise to repay.

Since short-term U.S. government bonds are for all intents and purposes completely risk-free and 100 percent likely to repay as promised, their rates of return are *purely* compensation for time preference and the fact that people must be compensated for delaying their own consumption opportunities when they lend money to the government. One consequence of this fact is that the rate of return earned by short-term U.S. government bonds is often referred to as the **risk-free interest rate,** or i^f, to clearly indicate that the rate of return that they generate is not in any way a compensation for risk.

It should be kept in mind, however, that the Federal Reserve has the power to change the risk-free interest rate generated by short-term U.S. government bonds. As discussed in Chapter 34, the Federal Reserve can use open-market operations to lower or raise the federal funds interest rate by making large purchases or sales of U.S. securities in the bond market. These open-market operations affect the money supply, which affects all interest rates, including the rates on short-term U.S. government bonds. This means that the Federal Reserve indirectly determines the risk-free interest rate and, consequently, the compensation that investors receive for being patient. As we will soon demonstrate, this fact is very important because by manipulating the reward for being patient, the Federal Reserve can affect the rate of return and prices of not only government bonds but all assets.

QUICK REVIEW 35.5

- The average expected rate of return is the probability-weighted average of an investment's possible future returns.
- Beta measures the nondiversifiable risk of an investment relative to the amount of nondiversifiable risk facing the market portfolio, which is the portfolio containing every asset available in the financial markets.
- Because investors dislike risk, riskier investments must offer higher rates of return to compensate investors for bearing more risk.
- Average expected rates of return compensate investors for both risk and time preference, which is the preference most people have to consume sooner rather than later.

The Security Market Line

LO35.8 Explain how the Security Market Line illustrates the compensation that investors receive for time preference and nondiversifiable risk and why arbitrage will tend to move all assets onto the Security Market Line.

Investors must be compensated for time preference as well as for the amount of nondiversifiable risk that an investment carries with it. This section introduces a simple model called the **Security Market Line,** which indicates how this compensation is determined for all assets no matter what their respective risk levels happen to be.

The underlying logic of the model is this: Any investment's average expected rate of return has to be the sum of two parts—one that compensates for time preference and another that compensates for risk. That is,

$$\begin{aligned}\text{Average expected rate of return} &= \text{rate that compensates for time preference} \\ &\quad + \text{rate that compensates for risk}\end{aligned}$$

As we explained, the compensation for time preference is equal to the risk-free interest rate, i^f, that is paid on short-term government bonds. As a result, this equation can be simplified to

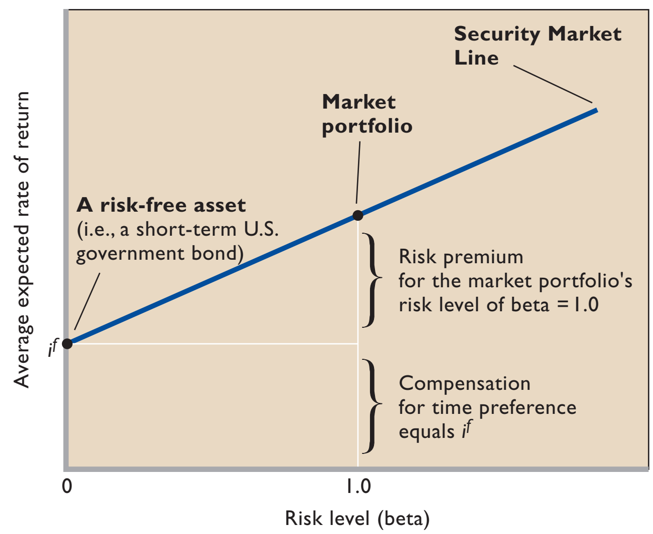
$$\text{Average expected rate of return} = i^f + \text{rate that compensates for risk}$$

Finally, because economists typically refer to the rate that compensates for risk as the **risk premium,** this equation can be simplified even further to

$$\text{Average expected rate of return} = i^f + \text{risk premium}$$

Naturally, the size of the risk premium that compensates for risk will vary depending on how risky an investment happens to be. In particular, it will depend on how big or small the investment's beta is. Investments with large betas and lots of nondiversifiable risk will require larger risk premiums than investments that have small betas and low levels of nondiversifiable risk. And, in the most extreme case, risk-free assets that have betas equal to zero will require no compensation for risk at all since they have no risk to compensate for.

FIGURE 35.1 The Security Market Line. The Security Market Line shows the relationship between average expected rates of return and risk levels that must hold for every asset or portfolio trading in the financial markets. Each investment's average expected rate of return is the sum of the risk-free interest rate that compensates for time preference as well as a risk premium that compensates for the investment's level of risk. The Security Market Line's upward slope reflects the fact that investors must be compensated for higher levels of risk with higher average expected rates of return.



This logic is translated into the graph presented in Figure 35.1. The horizontal axis of Figure 35.1 measures risk levels using beta; the vertical axis measures average expected rates of return. As a result, any investment can be plotted on Figure 35.1 just as long as we know its beta and its average expected rate of return. We have plotted two investments in Figure 35.1. The first is a risk-free short-term U.S. government bond, which is indicated by the lower-left dot in the figure. The second is the market portfolio, which is indicated by the upper-right dot in the figure.

The lower dot marking the position of the risk-free bond is located where it is because it is a risk-free asset having a beta $= 0$ and because its average expected rate of return is given by i^f. These values place the lower dot i^f percentage points up the vertical axis, as shown in Figure 35.1. Note that this location conveys the logic that because this asset has no risk, its average expected rate of return only has to compensate investors for time preference—which is why its average expected rate of return is equal to precisely i^f and no more.

The market portfolio, by contrast, is risky so that its average expected rate of return must compensate investors not only for time preference but also for the level of risk to which the market portfolio is exposed, which by definition is beta = 1.0. This implies that the vertical distance from the horizontal axis to the upper dot is equal to the sum of i^f and the market portfolio's risk premium.

The straight line connecting the risk-free asset's lower dot and the market portfolio's upper dot is called the Security Market Line, or SML. The SML is extremely important because it defines the relationship between average expected rates of return and risk levels that must hold for all assets and all portfolios trading in the financial markets. The SML illustrates the idea that every asset's average expected rate of return is the sum of a rate of return that compensates for time preference and a rate of return that compensates for risk. More specifically, the SML has a vertical intercept equal to the rate of interest earned by short-term U.S. government bonds and a positive slope that compensates investors for risk.

As we explained earlier, the precise location of the intercept at any given time is determined by the Federal Reserve's monetary policy and how it affects the rate of return on short-term U.S. government bonds. The slope of the SML, however, is determined by investors' feelings about risk and how much compensation they require for dealing with it. If investors greatly dislike risk, then the SML will have to be very steep, so that any given increase in risk on the horizontal axis will result in a very large increase in compensation as measured by average expected rates of return on the vertical axis. On the other hand, if investors dislike risk only moderately, then the SML will be relatively flat since any given increase in risk on the horizontal axis would require only a moderate increase in compensation as measured by average expected rates of return on the vertical axis.

It is important to realize that once investor preferences about risk have determined the slope of the SML and monetary policy has determined its vertical intercept, the SML plots out the precise relationship between risk levels and average expected rates of return *that should hold for every asset.* For instance, consider Figure 35.2, where there is an asset whose risk level on the horizontal axis is beta = X. The SML tells us that every asset with that risk level should have an average expected rate of return equal to Y on the vertical axis. This average expected rate of return exactly compensates for both time preference and the fact that the asset in question is exposed to a risk level of beta = X.

Finally, it should be pointed out that arbitrage will ensure that all investments having an identical level of risk also will have an identical rate of return—the return given by the SML. This is illustrated in Figure 35.3, where the three assets A, B, and C all share the same risk level of beta = X but initially have three different average expected rates of return. Since asset B lies on the SML, it has the average expected rate of return Y that precisely compensates investors for time preference and risk level X. Asset A, however, has a higher average expected rate of return that overcompensates investors while asset C has a lower average expected rate of return that undercompensates investors.

FIGURE 35.2 Risk levels determine average expected rates of return. The Security Market Line can be used to determine an investment's average expected rate of return based on its risk level. In this figure, investments having a risk level of beta = X will have an average expected rate of return of Y percent per year. This average expected rate of return will compensate investors for time preference in addition to providing them exactly the right sized risk premium to compensate them for dealing with a risk level of beta = X.

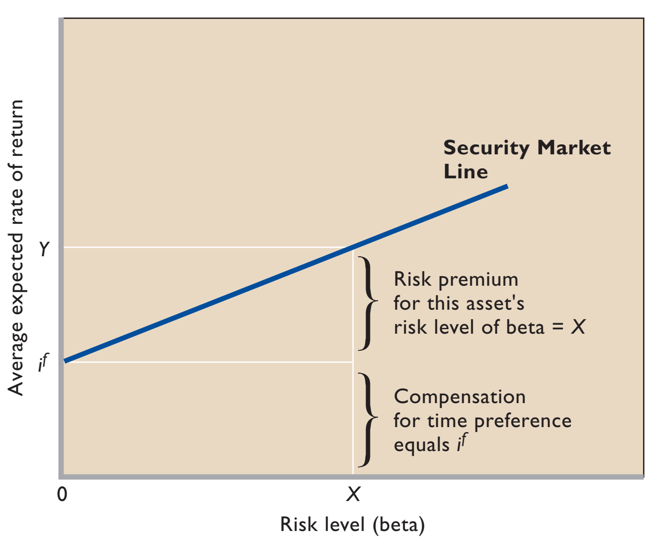
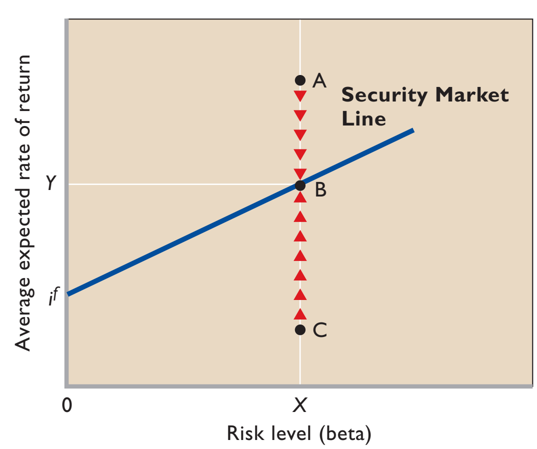


FIGURE 35.3 Arbitrage and the Security Market Line. Arbitrage pressures will tend to move any asset or portfolio that lies off the Security Market Line back onto the line. Investors will increase their purchases of asset A, driving up its price and decreasing its average expected rate of return. They will decrease their purchases of asset C, reducing its price and raising its return. Therefore, assets A, B, and C will all end up on the Security Market Line with each having the same average expected rate of return, Y, at risk level (beta) X. This average expected rate of return will fully compensate the investors for time preference plus nondiversifiable risk as measured by beta.



CONSIDER THIS . . .

Ponzi Schemes

Ponzi schemes—named after 1920s fraudster Charles Ponzi—are investments in which investors are unknowingly paid returns directly from the investments made by new investors. A successful Ponzi scheme attracts increasing amounts of money, much of which is siphoned off by the promoter. Common to all such schemes is the promise of high investment returns relative to investment risk. In Figure 35.1, the implied risk-return point lies someplace well above the SML.

Sophisticated investors know that arbitrage reduces returns, and even novice investors intuitively understand that good investment opportunities cannot continue unless they are exclusive. So promoters of Ponzi schemes concoct various reasons why such arbitrage does not occur. They claim there is a secret investment strategy or an exotic investment opportunity known only to the promoter.

Trust plays an important role in many Ponzi schemes. The promoter often attracts initial investors from affinity groups, targeting fellow church members or people who belong to the same country club. Current investors then become unwitting promoters of the Ponzi scheme because they recommend the attractive investment to family and friends.

The largest ever Ponzi scheme—run by Bernie Madoff—lasted for decades until it was exposed in 2008. On paper, investments with Madoff offered steady, solid returns year after year. Better still, those returns continued to accrue even during periods of financial crisis and recession. Madoff seemingly had found a strategy to eliminate nondiversifiable risk! But the true risk of investing with Madoff became distressingly apparent when the scheme collapsed. Madoff investors lost $13 billion of invested money and $65 billion of money they thought they had in their investment portfolios. Sadly, some individuals lost their entire life savings.

Arbitrage pressures will quickly eliminate these over- and undercompensations. For instance, consider what will happen to asset A. Investors will be hugely attracted to its overly high rate of return and will rush to buy it. That will drive up its price. But because average expected rates of return and prices are inversely related, the increase in price will cause its average expected rate of return to fall. Graphically, this means that asset A will move vertically downward as illustrated in Figure 35.3. And it will continue to move vertically downward until it reaches the SML since only then will it have the average expected rate of return Y that properly compensates investors for time preference and risk level X.

A similar process also will move asset C back to the SML. Investors will dislike the fact that its average expected rate of return is so low. This will cause them to sell it, driving down its price. Since average expected rates of return and prices are inversely related, this will cause its average expected rate of return to rise, thereby causing C to rise vertically as illustrated in Figure 35.3. And as with point A, point C will continue to rise until it reaches the SML, since only then will it have the average expected rate of return Y that properly compensates investors for time preference and risk level X.

Security Market Line: Applications

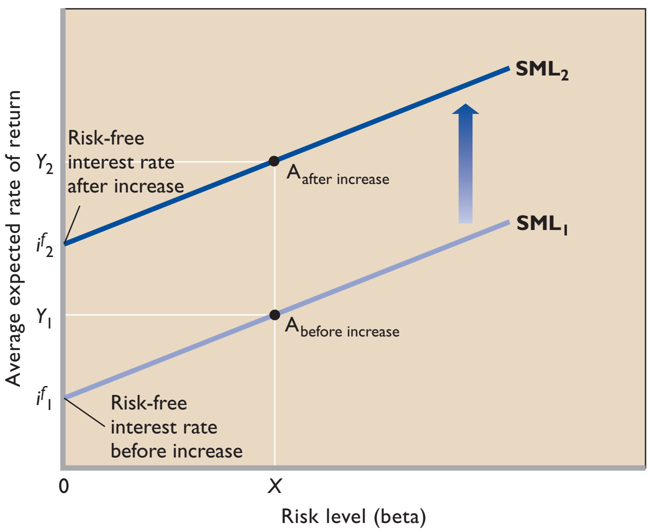
The SML analysis is highly useful in clarifying why investors scrutinize the intentions and actions of the Federal Reserve and change their behaviors during financial crises.

An Increase in the Risk-Free Rate by the Fed We have just explained how the position of the Security Market Line is fixed by two factors. The vertical intercept is set by the risk-free interest rate while the slope is determined by the amount of compensation investors demand for bearing nondiversifiable risk. As a result, changes in either one of these factors can shift the SML and thereby cause large changes in both average expected rates of return and asset prices.

As an example, consider what happens to the SML if the Federal Reserve changes policy and uses open-market operations (described in Chapter 34) to reduce the money supply, raise the federal funds rate, and increase other interest rates such as those on short-term U.S. government bonds. Since the risk-free interest rate earned by these bonds is also the SML's vertical intercept, an increase in their interest rate will move the SML's vertical intercept upward, as illustrated in Figure 35.4. The result is a parallel upward shift of the SML from SML_1 to SML_2. (The shift is parallel because nothing has happened that would affect the SML's slope, which is determined by the amount of compensation that investors demand for bearing risk.)

Notice what this upward shift implies. Not only does the rate of return on short-term U.S. government bonds increase when the Federal Reserve changes policy, but the rate of return on risky assets increases as well. For instance, consider asset A, which originally has rate of return Y_1. After the SML shifts upward, asset A ends up with the higher rate of return Y_2. There is a simple intuition behind this increase. Risky assets must compete with risk-free assets for investor money. When the Federal Reserve increases the rate of return on risk-free short-term U.S. government bonds, they become more attractive to investors. But to get the money to buy more risk-free bonds, investors have to sell risky assets. This drives down their

FIGURE 35.4 An increase in risk-free interest rates causes the SML to shift up vertically. The risk-free interest rate set by the Federal Reserve is the Security Market Line's vertical intercept. Consequently, if the Federal Reserve increases the risk-free interest rate, the Security Market Line's vertical intercept will move up. This rise in the risk-free interest rate will result in a decline in all asset prices and thus an increase in the average expected rate of return on all assets. So the Security Market Line will shift up parallel from SML_1 to SML_2. Here, asset A with risk level beta = X sees its average expected rate of return rise from Y_1 to Y_2.



prices and—because prices and average expected rates of return are inversely related—causes their average expected rates of return to increase. The result is that asset A moves up vertically in Figure 35.4, its average expected rate of return increasing from Y_1 to Y_2 as investors reallocate their wealth from risky assets like asset A to risk-free bonds.

This process explains why investors are so watchful of the Federal Reserve and keenly interested in its policies. Any increase in the risk-free interest rate leads to a decrease in asset prices that directly reduces investors' wealth.

The Fed's power to change asset prices through monetary policy stems entirely from the fact that changes in the risk-free rate shift the SML and thus totally redefine the investment opportunities available in the economy. As the set of options changes, investors modify their portfolios to obtain the best possible combination of risk and returns from the new set of investment options. In doing so, they engage in massive amounts of buying and selling to get rid of assets they no longer want and acquire assets that they now desire. These massive changes in supply and demand for financial assets are what cause their prices to change so drastically when the Federal Reserve alters the risk-free interest rate.

The Security Market Line during the Great Recession The special circumstances of the financial markets during the recession of 2007–2009 provide an excellent illustration of both the impact of Federal Reserve actions and the idea of *time-varying risk premium*. The latter is the reality that the premium demanded by investors to take on risk may vary from one period (and one set of economic circumstances) to another period (and a different set of economic circumstances).

The Federal Reserve used expansionary monetary policy during this period to lower interest rates, including the interest rates of short-term U.S. government bonds. Because the risk-free interest rate earned by these securities locates the vertical intercept of the Securities Market Line, the actual SML for the economy shifted downward from that shown in Figure 35.1. This decline would be portrayed as the opposite of the upward shift that we illustrate in Figure 35.4.

But wouldn't we expect stock market prices to rise when the risk-free rate of return falls? That certainly did *not* happen in 2007 and 2008. Yes, normally, stock market prices rise when the risk-free interest rate falls. But during this unusual period, investors became very fearful about losses from investments in general and began to look for any place of financial safety. As their appetite for risk decreased, they demanded a much higher rate of compensation for taking on any particular level of risk. In terms of Figure 35.4, the slope of the SML greatly increased. Thus, between the Fed's deliberate reduction of the risk-free rate and investors' diminished appetite for risk, two things happened at once to the SML: (1) its intercept (the risk-free rate) dramatically fell and (2) the SML became much steeper. In Figure 35.1, these two effects would be shown by a much steeper SML emanating from a much lower point on the vertical axis.

The increase in the slope of the SML overwhelmed the decline in the intercept. Investors sold off stocks, which greatly reduced stock prices, even though the risk-free interest rate fell.

QUICK REVIEW 35.6

- The Security Market Line (SML) is a straight, upsloping line showing how the average expected rates of return on investments vary with their respective levels of nondiversifiable risk as measured by beta.
- Arbitrage ensures that every asset in the economy should plot onto the SML.
- The positive slope of the SML reflects the fact that investors dislike nondiversifiable risk; the steeper the slope, the greater the dislike.
- The Fed can shift the entire SML upward or downward by using monetary policy to change the risk-free interest rate on short-term U.S. bonds.

LAST WORD

Index Funds versus Actively Managed Funds

Do Actively Managed Funds Outperform Passively Managed Index Funds That Have the Same Risk?

Mutual fund investors have a choice between putting their money into actively managed mutual funds or into passively managed index funds. Actively managed funds constantly buy and sell assets in an attempt to build portfolios that will generate average expected rates of return that are higher than those of other portfolios possessing a similar level of risk. In terms of Figure 35.3, they try to construct portfolios similar to point A, which has the same level of risk as portfolio B but a much higher average expected rate of return. By contrast, the portfolios of index funds simply mimic the assets that are included in their underlying indexes and make no attempt whatsoever to generate higher returns than other portfolios having similar levels of risk.

As a result, expecting actively managed funds to generate higher rates of return than index funds would seem only natural. Surprisingly, however, the exact opposite actually holds true. Once costs are taken into account, the average returns generated by index funds trounce those generated by actively managed funds by well over 1 percent per year. Now, 1 percent per year may not sound like a lot, but the compound interest formula of equation 1 shows that $10,000 growing for 30 years at 10 percent per year becomes $170,449.40, whereas that same amount of money growing at 11 percent for 30 years becomes $220,892.30. For anyone saving for retirement, an extra 1 percent per year is a very big deal.

Why do actively managed funds do so much worse than index funds? The answer is twofold. First, arbitrage makes it virtually impossible for actively managed funds to select portfolios that will do any better than index funds that have similar levels of risk. As a result, *before taking costs into account*, actively managed funds and index funds produce very similar returns. Second, actively managed funds charge their investors much higher fees than do passively managed funds, so that, *after taking costs into account*, actively managed funds do worse by about 1 percent per year.

Let us discuss each of these factors in more detail. The reason that actively managed funds cannot do better than index funds before taking costs into account has to do with the power of arbitrage to ensure that investments having equal levels of risk also have equal average expected rates of return. As we explained above with respect to Figure 35.3, assets and portfolios that deviate from the Security Market Line (SML) are very quickly forced back onto the SML by arbitrage, so that assets and portfolios with equal levels of risk have equal average expected rates of return. This implies that index funds and actively managed funds with equal levels of risk will end up with identical average expected rates of return despite the best efforts of actively managed funds to produce superior returns.

The reason actively managed funds charge much higher fees than index funds is because they run up much higher costs while trying to produce superior returns. Not only do they have to pay large salaries to professional fund managers, but they also have to pay for the massive amounts of trading that those managers engage in as they buy and sell assets in their quest to produce superior returns. The costs of running an index fund are, by contrast, very small since changes are made to an index fund's portfolio only on the rare occasions when the fund's underlying index changes. As a result, trading costs are low and there is no need to pay for a professional manager. The overall result is that while the largest and most popular index fund currently charges its investors only 0.18 percent per year for its services, the typical actively managed fund charges more than 1.5 percent per year.

So why are actively managed funds still in business? The answer may well be that index funds are boring. Because they are set up to mimic indexes that are in turn designed to show what average performance levels are, index funds are by definition stuck with average rates of return and absolutely no chance to exceed average rates of return. For investors who want to try to beat the average, actively managed funds are the only way to go.

SUMMARY

LO35.1 Define *financial economics* and distinguish between economic investment and financial investment.

Financial investment focuses its attention on investor preferences and how they affect the trading and pricing of the wide variety of financial assets available in the modern economy, including stocks, bonds, and real estate. Remember to distinguish between economic investment (paying for additions to the capital stock) and financial investment (buying an existing asset or building a new asset in the expectation of financial gain).

LO35.2 Explain the time value of money and how compound interest can be used to calculate the present value of any future amount of money.

The compound interest formula states that if X_0 dollars is invested today at interest rate i and allowed to grow for t years, it will become $(1 + i)^t X_0$ dollars in t years. The present value formula rearranges the compound interest formula. It tells investors the current number of dollars that they would have to invest today to receive X_t dollars in t years.

All financial assets available to investors have a common characteristic: In exchange for a certain price today, they all promise to make one or more payments in the future. A risk-free investment's proper current price is simply equal to the sum of the present values of each of the investment's expected future payments.

LO35.3 Identify and distinguish between the most common financial investments: stocks, bonds, and mutual funds.

Stocks give shareholders the right to share in any future profits that corporations may generate. The main risk of stock investing is that future profits are unpredictable and that companies may go bankrupt. Bonds provide bondholders the right to receive a fixed stream of future payments that serve to repay the loan. Bonds are risky because of the possibility that the corporations or government bodies that issued the bonds may default on them, or make less than the promised payments. Mutual funds own and manage portfolios of bonds and stocks; fund investors get the returns generated by those portfolios. The risks of mutual funds to investors reflect the risks of the stocks and bonds that are in their respective portfolios.

LO35.4 Relate how percentage rates of return provide a common framework for comparing assets and explain why asset prices and rates of return are inversely related.

Investors evaluate the possible future returns to risky investments using average expected rates of return, which give higher weight to outcomes that are more likely to happen. Average expected rates of return are inversely related to an asset's current price.

LO35.5 Define and utilize the concept of arbitrage.

Arbitrage is the process whereby investors equalize the average expected rates of return generated by identical or nearly identical assets. If two identical assets have different rates of return, investors will sell the asset with the lower rate of return and buy the asset with the higher rate of return. As investors buy the asset with the higher rate of return, its price will rise, reducing its average expected rate of return. At the same time, as investors sell the asset with the lower rate of return, its price will fall, raising its average expected rate of return. The process will continue until the average expected rates of return are equal on the two investments.

LO35.6 Describe how the word *risk* is used in financial economics and explain the difference between diversifiable and nondiversifiable risk.

In finance, an asset is risky if its future payments are uncertain. What matters is not whether the payments are big or small, positive or negative, or good or bad—only that they are not guaranteed ahead of time. Risks that can be canceled out by diversification are called diversifiable risks. Risks that cannot be canceled out by diversification are called nondiversifiable risks.

LO35.7 Convey why investment decisions are determined primarily by investment returns and nondiversifiable risk and how investment returns compensate for being patient and for bearing nondiversifiable risk.

Beta is a statistic that measures the nondiversifiable risk of an asset or portfolio relative to the amount of nondiversifiable risk facing the market portfolio. Because the market portfolio contains every asset trading in the financial markets, it is completely diversified and therefore exposed to only nondiversifiable risk. By tradition, its beta is set at 1.0. Thus, an investment that has a beta of 0.5 is exposed to half as much nondiversifiable risk as the market portfolio.

Investors dislike risk and therefore demand compensation for being exposed to it. This compensation takes the form of higher average expected rates of return. The riskier the asset, the greater is the average expected rate of return. Because a well-diversified portfolio has no diversifiable risk, investors will need to be compensated only for the asset's level of nondiversifiable risk as measured by beta.

Average expected rates of return also must compensate for time preference and the fact that, other things equal, people prefer to consume sooner rather than later. Consequently, an asset's average expected rate of return will be the sum of the rate of return that compensates for time preference plus the rate of return that compensates for the asset's level of nondiversifiable risk as measured by beta.

LO35.8 Explain how the Security Market Line illustrates the compensation that investors receive for time preference and nondiversifiable risk and why arbitrage will tend to move all assets onto the Security Market Line.

The rate of return that compensates for time preference is assumed to be equal to the rate of interest generated by short-term U.S. government bonds. These bonds are considered to be risk-free, meaning that their rate of return must purely be compensation for time preference and not nondiversifiable risk.

The Security Market Line (SML) is a straight upsloping line showing how the average expected rates of return on assets and portfolios in the economy vary with their respective levels of nondiversifiable risk as measured by beta. Arbitrage ensures that every asset in the economy should plot onto the SML. The slope of the SML reflects the investors' dislike for nondiversifiable risk, with steeper slopes reflecting greater dislike for that risk.

The Fed can shift the entire SML upward or downward by using monetary policy to change the risk-free interest rate on short-term U.S. bonds. When the SML shifts, the average expected rate of return on all assets changes. Because average expected rates of return are inversely related to asset prices, the shift in the SML also will change asset prices. Therefore, the Federal Reserve's ability to change short-run interest rates also enables it to change asset prices throughout the economy.

TERMS AND CONCEPTS

economic investment
financial investment
present value
compound interest
stocks
bankrupt
limited liability rule
capital gains
dividends
bonds
default
mutual funds
portfolios
index funds
actively managed funds
passively managed funds
percentage rate of return
arbitrage
risk
diversification
diversifiable risk
nondiversifiable risk
average expected rate of return
probability-weighted average
beta
market portfolio
time preference
risk-free interest rate
Security Market Line
risk premium

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Suppose that the city of New York issues bonds to raise money to pay for a new tunnel linking New Jersey and Manhattan. An investor named Susan buys one of the bonds on the same day that the city of New York pays a contractor for completing the first stage of construction. Is Susan making an economic or a financial investment? What about the city of New York? **LO35.1**
2. What is compound interest? How does it relate to the formula $X_t = (1 + i)^t X_0$? What is present value? How does it relate to the formula $X_t/(1 + i)^t = X_0$? **LO35.2**
3. How do stocks and bonds differ in terms of the future payments that they are expected to make? Which type of investment (stocks or bonds) is considered to be more risky? Given what you know, which investment (stocks or bonds) do you think commonly goes by the nickname "fixed income"? **LO35.3**
4. What are mutual funds? What different types of mutual funds are there? And why do you think they are so popular with investors? **LO35.3**
5. Corporations often distribute profits to their shareholders in the form of dividends, which are simply checks mailed out to shareholders. Suppose that you have the chance to buy a share in a fashion company called Rogue Designs for \$35 and that the company will pay dividends of \$2 per year on that share every year. What is the annual percentage rate of return? Next, suppose that you and other investors could get a 12 percent per year rate of return by owning the stocks of other very similar fashion companies. If investors care only about rates of return, what should happen to the share price of Rogue Designs? (Hint: This is an arbitrage situation.) **LO35.5**
6. Why is it reasonable to ignore diversifiable risk and care only about nondiversifiable risk? What about investors who put all their money into only a single risky stock? Can they properly ignore diversifiable risk? **LO35.6**
7. If we compare the betas of various investment opportunities, why do the assets that have higher betas also have higher average expected rates of return? **LO35.7**
8. In this chapter we discussed short-term U.S. government bonds. But the U.S. government also issues longer-term bonds with horizons of up to 30 years. Why do 20-year bonds issued by the U.S. government have lower rates of return than 20-year bonds issued by corporations? And which would you consider more likely, that longer-term U.S. government

bonds have a higher interest rate than short-term U.S. government bonds, or vice versa? Explain. **LO35.7**

9. What determines the vertical intercept of the Security Market Line (SML)? What determines its slope? And what will happen to an asset's price if it initially plots onto a point above the SML? **LO35.8**
10. Suppose that the Federal Reserve thinks that a stock market bubble is occurring and wants to reduce stock prices. What should it do to interest rates? **LO35.8**
11. Consider another situation involving the SML. Suppose that the risk-free interest rate stays the same, but that investors' dislike of risk grows more intense. Given this change, will average expected rates of return rise or fall? Next, compare what will happen to the rates of return on low-risk and high-risk investments. Which will have a larger increase in average expected rates of return, investments with high betas or investments with low betas? And will high-beta or low-beta investments show larger percentage changes in their prices? **LO35.8**
12. **LAST WORD** Why is it so hard for actively managed funds to generate higher rates of return than passively managed index funds having similar levels of risk? Is there a simple way for an actively managed fund to increase its average expected rate of return?

REVIEW QUESTIONS

1. Identify each of the following investments as either an economic investment or a financial investment. **LO35.1**
   a. A company builds a new factory.
   b. A pension plan buys some Google stock.
   c. A mining company sets up a new gold mine.
   d. A woman buys a 100-year-old farmhouse in the countryside.
   e. A man buys a newly built home in the city.
   f. A company buys an old factory.
2. It is a fact that $(1 + 0.12)^3 = 1.40$. Knowing that to be true, what is the present value of $140 received in three years if the annual interest rate is 12 percent? **LO35.2**
   a. $1.40.
   b. $12.
   c. $100.
   d. $112.
3. Asset X is expected to deliver 3 future payments. They have present values of, respectively, $1,000, $2,000, and $7,000. Asset Y is expected to deliver 10 future payments, each having a present value of $1,000. Which of the following statements correctly describes the relationship between the current price of Asset X and the current price of Asset Y? **LO35.3**
   a. Asset X and Asset Y should have the same current price.
   b. Asset X should have a higher current price than Asset Y.
   c. Asset X should have a lower current price than Asset Y.
4. Tammy can buy an asset this year for $1,000. She is expecting to sell it next year for $1,050. What is the asset's anticipated percentage rate of return? **LO35.4**
   a. 0 percent.
   b. 5 percent.
   c. 10 percent.
   d. 15 percent.
5. Sammy buys stock in a suntan-lotion maker and also stock in an umbrella maker. One stock does well when the weather is good; the other does well when the weather is bad. Sammy's portfolio indicates that "weather risk" is a ____________ risk. **LO35.6**
   a. Diversifiable.
   b. Nondiversifiable.
   c. Automatic.
6. An investment has a 50 percent chance of generating a 10 percent return and a 50 percent chance of generating a 16 percent return. What is the investment's average expected rate of return? **LO35.7**
   a. 10 percent.
   b. 11 percent.
   c. 12 percent.
   d. 13 percent.
   e. 14 percent.
   f. 15 percent.
   g. 16 percent.
7. If an investment has 35 percent more nondiversifiable risk than the market portfolio, its beta will be: **LO35.7**
   a. 35.
   b. 1.35.
   c. 0.35.
8. The interest rate on short-term U.S. government bonds is 4 percent. The risk premium for any asset with a beta = 1.0 is 6 percent. What is the average expected rate of return on the market portfolio? **LO35.7**
   a. 0 percent.
   b. 4 percent.
   c. 6 percent.
   d. 10 percent.
9. Suppose that an SML indicates that assets with a beta = 1.15 should have an average expected rate of return of 12 percent per year. If a particular stock with a beta = 1.15 currently has an average expected rate of return of 15 percent, what should we expect to happen to its price? **LO35.8**
   a. Rise.
   b. Fall.
   c. Stay the same.
10. If the Fed increases interest rates, the SML will shift ____________ and asset prices will ____________. **LO35.8**
    a. Down; rise.
    b. Down; fall.
    c. Up; rise.
    d. Up; fall.

PROBLEMS

1. Suppose that you invest $100 today in a risk-free investment and let the 4 percent annual interest rate compound. Rounded to full dollars, what will be the value of your investment 4 years from now? **LO35.2**
2. Suppose that you desire to get a lump-sum payment of $100,000 two years from now. Rounded to full dollars, how many current dollars will you have to invest today at 10 percent interest to accomplish your goal? **LO35.2**
3. Suppose that a risk-free investment will make three future payments of $100 in one year, $100 in two years, and $100 in three years. If the Federal Reserve has set the risk-free interest rate at 8 percent, what is the proper current price of this investment? What is the price of this investment if the Federal Reserve raises the risk-free interest rate to 10 percent? **LO35.2**
4. Consider an asset that costs $120 today. You are going to hold it for 1 year and then sell it. Suppose that there is a 25 percent chance that it will be worth $100 in a year, a 25 percent chance that it will be worth $115 in a year, and a 50 percent chance that it will be worth $140 in a year. What is its average expected rate of return? Next, figure out what the investment's average expected rate of return would be if its current price were $130 today. Does the increase in the current price increase or decrease the asset's average expected rate of return? At what price would the asset have a zero average expected rate of return? **LO35.4**
5. Suppose initially that two assets, A and B, will each make a single guaranteed payment of $100 in 1 year. But asset A has a current price of $80 while asset B has a current price of $90. **LO35.6**
   - **a.** What are the rates of return of assets A and B at their current prices? Given these rates of return, which asset should investors buy and which asset should they sell?
   - **b.** Assume that arbitrage continues until A and B have the same expected rate of return. When arbitrage ends, will A and B have the same price?

   Next, consider another pair of assets, C and D. Asset C will make a single payment of $150 in one year, while D will make a single payment of $200 in one year. Assume that the current price of C is $120 and that the current price of D is $180.
   - **c.** What are the rates of return of assets C and D at their current prices? Given these rates of return, which asset should investors buy and which asset should they sell?
   - **d.** Assume that arbitrage continues until C and D have the same expected rate of return. When arbitrage ends, will C and D have the same price?

   Compare your answers to questions *a* through *d* before answering question *e*.
   - **e.** We know that arbitrage will equalize rates of return. Does it also guarantee to equalize prices? In what situations will it equalize prices?
6. **ADVANCED ANALYSIS** Suppose that the equation for the SML is $Y = 0.05 + 0.04X$, where Y is the average expected rate of return, 0.05 is the vertical intercept, 0.04 is the slope, and X is the risk level as measured by beta. What is the risk-free interest rate for this SML? What is the average expected rate of return at a beta of 1.5? What is the value of beta at an average expected rate of return of 7 percent? **LO35.8**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

PART **TEN**

EXTENSIONS AND ISSUES

CHAPTER **36**

Extending the Analysis of Aggregate Supply

Learning Objectives

LO36.1 Explain the relationship between short-run aggregate supply and long-run aggregate supply.

LO36.2 Discuss how to apply the "extended" (short-run/long-run) AD-AS model to inflation, recessions, and economic growth.

LO36.3 Explain the short-run trade-off between inflation and unemployment (the Phillips Curve).

LO36.4 Discuss why there is no long-run trade-off between inflation and unemployment.

LO36.5 Explain the relationship between tax rates, tax revenues, and aggregate supply.

During the early years of the Great Depression, many economists suggested that the economy would correct itself in the *long run* without government intervention. To this line of thinking, economist John Maynard Keynes remarked, "In the long run we are all dead!"

For several decades following the Great Depression, macroeconomists understandably focused on refining fiscal policy and monetary policy to smooth business cycles and address the problems of unemployment and inflation. The main emphasis was on short-run problems and policies associated with the business cycle.

But over people's lifetimes, and from generation to generation, the long run is tremendously important for economic well-being. For that reason, macroeconomists have refocused attention on long-run macroeconomic adjustments, processes, and outcomes. The renewed emphasis on the long run has produced significant insights about aggregate supply, economic growth, and economic development. We will also see in the next chapter that it has renewed historical debates over the causes of macro instability and the effectiveness of stabilization policy.

Our goals in this chapter are to extend the analysis of aggregate supply to the long run, examine the inflation-unemployment relationship, and evaluate the effect of taxes on aggregate supply. The latter is a key concern of so-called *supply-side economics*.

From Short Run to Long Run

LO36.1 Explain the relationship between short-run aggregate supply and long-run aggregate supply.

In Chapter 30, we noted that in macroeconomics the difference between the **short run** and the **long run** has to do with the flexibility of input prices. Input prices are inflexible or even totally fixed in the short run but fully flexible in the long run. (By contrast, output prices are assumed under these definitions to be fully flexible in both the short run and the long run.)

The assumption that input prices are flexible only in the long run leads to large differences in the shape and position of the short-run aggregate supply curve and the long-run aggregate supply curve. As explained in Chapter 30, the short-run aggregate supply curve is an upsloping line, whereas the long-run aggregate supply curve is a vertical line situated directly above the economy's full-employment output level, Q_f.

We will begin this chapter by discussing how aggregate supply transitions *from* the short run *to* the long run. Once that is done, we will combine the long-run and short-run aggregate supply curves with the aggregate demand curve to form a single model that can provide insights into how the economy adjusts to economic shocks as well as changes in monetary and fiscal policy in both the short run and the long run. That will lead us to discuss how economic growth relates to long-run aggregate supply and how inflation and aggregate supply are related in the long run and the short run. We will conclude with a discussion of a particular set of economic policies that may help increase both short-run aggregate supply and long-run aggregate supply.

Short-Run Aggregate Supply

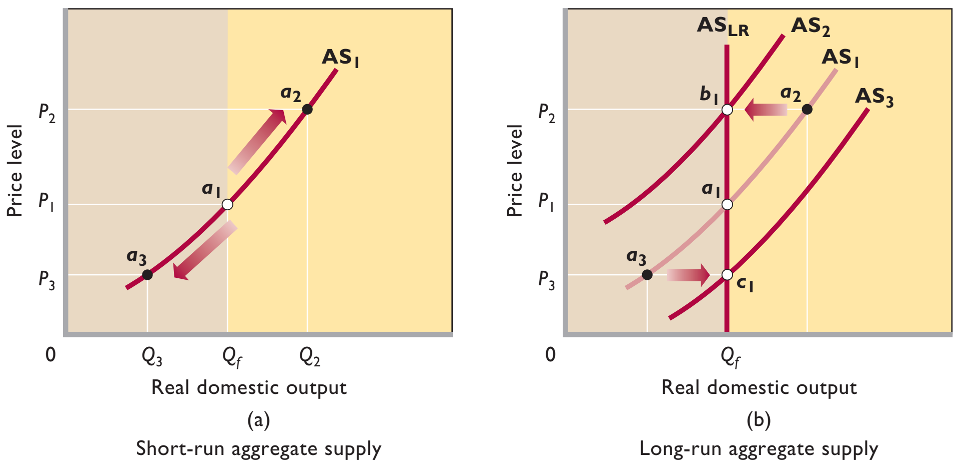
Our immediate objective is to demonstrate the relationship between short-run aggregate supply and long-run aggregate supply. We begin by briefly reviewing short-run aggregate supply.

Consider the short-run aggregate supply curve AS_1 in Figure 36.1a. This curve is based on three assumptions: (1) The initial price level is P_1, (2) firms and workers have established nominal wages on the expectation that this price level will persist, and (3) the price level is flexible both upward and downward. Observe from point a_1 that at price level P_1 the economy is operating at its full-employment output Q_f. This output is the real production forthcoming when the economy is operating at its natural rate of unemployment (or potential output).

Now let's review the short-run effects of changes in the price level, say, from P_1 to P_2 in Figure 36.1a. The higher prices associated with price level P_2 increase firms' revenues, and because their nominal wages and other input prices remain unchanged, their profits rise. Those higher profits lead firms to increase their output from Q_f to Q_2, and the economy moves from a_1 to a_2 on aggregate supply AS_1. At output Q_2 the economy is operating beyond its full-employment output. The firms make this possible by extending the work hours of part-time and full-time workers, enticing new workers such as homemakers and retirees into the labor force, and hiring and training the structurally unemployed. Thus, the nation's unemployment rate declines below its natural rate.

How will the firms respond when the price level *falls*, say, from P_1 to P_3 in Figure 36.1a? Because the prices they receive for their products are lower while the nominal wages they pay workers remain unchanged, firms discover that their revenues and profits have diminished or disappeared. So they reduce their production and employment, and, as shown by the movement from a_1 to a_3, real output falls to Q_3. Increased unemployment and a higher unemployment rate accompany the decline in real output. At output Q_3 the unemployment rate is greater than the natural rate of unemployment associated with output Q_f.

FIGURE 36.1 Short-run and long-run aggregate supply. (a) In the short run, nominal wages and other input prices do not respond to price-level changes and are based on the expectation that price level P_1 will continue. An increase in the price level from P_1 to P_2 increases profits and output, moving the economy from a_1 to a_2; a decrease in the price level from P_1 to P_3 reduces profits and real output, moving the economy from a_1 to a_3. The short-run aggregate supply curve therefore slopes upward. (b) In the long run, a rise in the price level results in higher nominal wages and other input prices and thus shifts the short-run aggregate supply curve to the left. Conversely, a decrease in the price level reduces nominal wages and shifts the short-run aggregate supply curve to the right. After such adjustments, the economy obtains equilibrium of points such as b_1 and c_1. Thus, the long-run aggregate supply curve is vertical at the full-employment output.



Long-Run Aggregate Supply

The outcomes are different in the long run. To see why, we need to extend the analysis of aggregate supply to account for changes in nominal wages that occur in response to changes in the price level. That will enable us to derive the economy's long-run aggregate supply curve.

We illustrate the implications for aggregate supply in Figure 36.1b. Again, suppose that the economy is initially at point a_1 (P_1 and Q_f). As we just demonstrated, an increase in the price level from P_1 to P_2 will move the economy from point a_1 to a_2 along the short-run aggregate supply curve AS_1. At a_2, the economy is producing at more than its potential output. This implies very high demand for productive inputs, so that input prices will begin to rise. In particular, the high demand for labor will drive up nominal wages, which will increase per unit production costs. As a result, the short-run supply curve shifts leftward from AS_1 to AS_2, which now reflects the higher price level P_2 and the new expectation that P_2, not P_1, will continue. The leftward shift in the short-run aggregate supply curve to AS_2 moves the economy from a_2 to b_1. Real output falls back to its full-employment level Q_f, and the unemployment rate rises to its natural rate.

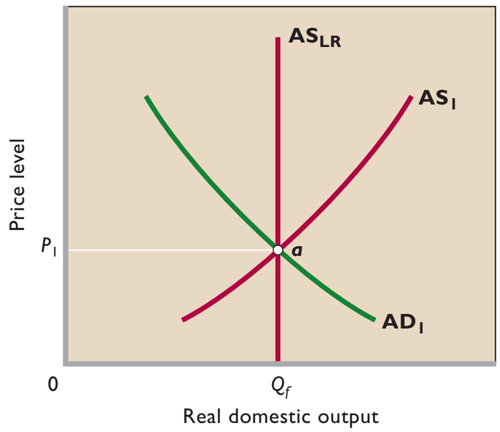
What is the long-run outcome of a *decrease* in the price level? Assuming eventual downward wage flexibility, a decline in the price level from P_1 to P_3 in Figure 36.1b works in the opposite way from a price-level increase. At first the economy moves from point a_1 to a_3 on AS_1. Profits are squeezed or eliminated because prices have fallen and nominal wages have not. But this movement along AS_1 is the short-run supply response that results only while input prices remain constant. As time passes, input prices will begin to fall because the economy is producing at below its full-employment output level. With so little output being produced, the demand for inputs will be low and their prices will begin to decline. In particular, the low demand for labor will drive down nominal wages and reduce per-unit production costs. Lower nominal wages therefore shift the short-run aggregate supply curve rightward from AS_1 to AS_3, and real output returns to its full-employment level of Q_f at point c_1.

By tracing a line between the long-run equilibrium points b_1, a_1, and c_1, we obtain a long-run aggregate supply curve. Observe that it is vertical at the full-employment level of real GDP. After long-run adjustments in nominal wages and other nominal input prices, real output is Q_f regardless of the specific price level.

Long-Run Equilibrium in the AD-AS Model

Figure 36.2 helps us understand the long-run equilibrium in the AD-AS model, now extended to include the

FIGURE 36.2 Equilibrium in the long-run AD-AS model. The long-run equilibrium price level P_1 and level of real output Q_f occur at the intersection of the aggregate demand curve AD_1, the long-run aggregate supply curve AS_{LR}, and the short-run aggregate supply curve AS_1.



distinction between short-run and long-run aggregate supply. (Hereafter, we will refer to this model as the extended AD-AS model, with "extended" referring to the inclusion of both the short-run and the long-run aggregate supply curves.)

In the short run, equilibrium occurs wherever the downsloping aggregate demand curve and upsloping short-run aggregate supply curve intersect. This can be at any level of output, not simply the full-employment level. Either a negative GDP gap or a positive GDP gap is possible in the short run.

But in the long run, the short-run aggregate supply curve adjusts as we just described. After those adjustments, long-run equilibrium occurs where the aggregate demand curve, vertical long-run aggregate supply curve, and short-run aggregate supply curve all intersect. Figure 36.2 shows the long-run outcome. Equilibrium occurs at point *a*, where AD_1 intersects both AS_{LR} and AS_1, and the economy achieves its full-employment (or potential) output, Q_f. At long-run equilibrium price level P_1 and output level Q_f, neither a negative GDP gap nor a positive GDP gap occurs. The economy's *natural rate of unemployment* prevails, meaning that the economy achieves full employment.

In the United States, output Q_f in Figure 36.2 implies a 4 to 5 percent unemployment rate. The natural rate of unemployment can vary from one time period to another and can differ between countries. But whatever the rate happens to be, it defines the level of potential output and establishes the location of the long-run AS curve.

QUICK REVIEW 36.1

- The short-run aggregate supply curve slopes upward because nominal wages and other input prices are fixed while output prices change.
- The long-run aggregate supply curve is vertical because input prices eventually rise in response to changes in output prices.
- The long-run equilibrium GDP and price level occur at the intersection of the aggregate demand curve, the long-run aggregate supply curve, and the short-run aggregate supply curve.
- In long-run equilibrium, the economy achieves its natural rate of unemployment and its full-potential real output.

Applying the Extended AD-AS Model

LO36.2 Explain how to apply the "extended" (short-run/long-run) AD-AS model to inflation, recessions, and economic growth.

The extended AD-AS model helps clarify the long-run aspects of demand-pull inflation, cost-push inflation, and recession.

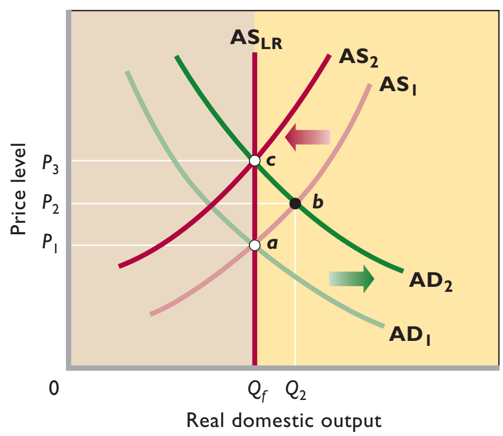
Demand-Pull Inflation in the Extended AD-AS Model

Recall that demand-pull inflation occurs when an increase in aggregate demand pulls up the price level. Earlier, we depicted this inflation by shifting an aggregate demand curve rightward along a stable aggregate supply curve (see Figure 30.8).

In our more complex version of aggregate supply, an increase in the price level eventually leads to an increase in nominal wages and thus a leftward shift of the short-run aggregate supply curve. This is shown in Figure 36.3, where we initially suppose the price level is P_1 at the intersection of aggregate demand curve AD_1, short-run supply curve AS_1, and long-run aggregate supply curve AS_{LR}. Observe that the economy is achieving its full-employment real output Q_f at point *a*.

Now consider the effects of an increase in aggregate demand as represented by the rightward shift from AD_1 to AD_2. This shift might result from any one of a number of factors, including an increase in investment spending or a rise in net exports. Whatever its cause, the increase in aggregate demand boosts the price level from P_1 to P_2 and expands real output from Q_f to Q_2 at point *b*. There, a positive GDP gap of $Q_2 - Q_f$ occurs.

FIGURE 36.3 Demand-pull inflation in the extended AD-AS model. An increase in aggregate demand from AD_1 to AD_2 drives up the price level and increases real output in the short run. But in the long run, nominal wages rise and the short-run aggregate supply curve shifts leftward, as from AS_1 to AS_2. Real output then returns to its prior level, and the price level rises even more. In this scenario, the economy moves from *a* to *b* and then eventually to *c*.



So far, none of this is new to you. But now the distinction between short-run aggregate supply and long-run aggregate supply becomes important. With the economy producing above potential output, inputs will be in high demand. Input prices including nominal wages will rise. As they do, the short-run aggregate supply curve will ultimately shift leftward such that it intersects long-run aggregate supply at point *c*.[1] There, the economy has reestablished long-run equilibrium, with the price level and real output now P_3 and Q_f, respectively. Only at point *c* does the new aggregate demand curve AD_2 intersect both the short-run aggregate supply curve AS_2 and the long-run aggregate supply curve AS_{LR}.

In the short run, demand-pull inflation drives up the price level and increases real output; in the long run, only the price level rises. In the long run, the initial increase in aggregate demand moves the economy along its vertical aggregate supply curve AS_{LR}. For a while, an economy can operate beyond its full-employment level of output. But the demand-pull inflation eventually causes adjustments of nominal wages that return the economy to its full-employment output Q_f.

[1]We say "ultimately" because the initial leftward shift in short-run aggregate supply will intersect the long-run aggregate supply curve AS_{LR} at price level P_2 (review Figure 36.1b). But the intersection of AD_2 and this new short-run aggregate supply curve (that is not shown in Figure 36.3) will produce a price level above P_2. (You may want to pencil this in to make sure that you understand this point.) Again nominal wages will rise, shifting the short-run aggregate supply curve farther leftward. The process will continue until the economy moves to point *c*, where the short-run aggregate supply curve is AS_2, the price level is P_3, and real output is Q_f.

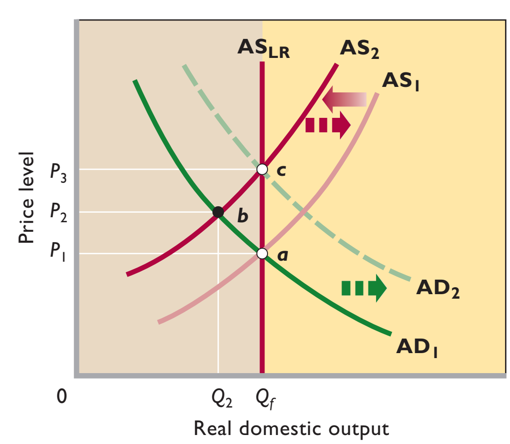
Cost-Push Inflation in the Extended AD-AS Model

Cost-push inflation arises from factors that increase the cost of production at each price level, shifting the aggregate supply curve leftward and raising the equilibrium price level. Previously (Figure 30.10), we considered cost-push inflation using only the short-run aggregate supply curve. Now we want to analyze that type of inflation in its long-run context.

Analysis Look at Figure 36.4, in which we again assume that the economy is initially operating at price level P_1 and output level Q_f (point *a*). Suppose that international oil producers agree to reduce the supply of oil to boost its price by, say, 100 percent. As a result, the per-unit production cost of producing and transporting goods and services rises substantially in the economy represented by Figure 36.4. This increase in per-unit production costs shifts the short-run aggregate supply curve to the left, as from AS_1 to AS_2, and the price level rises from P_1 to P_2 (as seen by comparing points *a* and *b*). In this case, the leftward shift of the aggregate supply curve is *not a response* to a price-level increase, as it was in our previous discussions of demand-pull inflation; it is the *initiating cause* of the price-level increase.

Policy Dilemma Cost-push inflation creates a dilemma for policymakers. Without some expansionary stabilization policy, aggregate demand in Figure 36.4 remains in

FIGURE 36.4 Cost-push inflation in the extended AD-AS model. Cost-push inflation occurs when the short-run aggregate supply curve shifts leftward, as from AS_1 to AS_2. If government counters the decline in real output by increasing aggregate demand to the broken line, the price level rises even more. That is, the economy moves in steps from *a* to *b* to *c*. In contrast, if government allows a recession to occur, nominal wages eventually fall and the aggregate supply curve shifts back rightward to its original location. The economy moves from *a* to *b* and eventually back to *a*.



place at AD_1 and real output declines from Q_f to Q_2. Government can counter this recession, negative GDP gap, and attendant high unemployment by using fiscal policy and monetary policy to increase aggregate demand to AD_2. But there is a potential policy trap here: An increase in aggregate demand to AD_2 will further raise inflation by increasing the price level from P_2 to P_3 (a move from point *b* to point *c*).

Suppose the government recognizes this policy trap and decides not to increase aggregate demand from AD_1 to AD_2 (you can now disregard the dashed AD_2 curve) and instead decides to allow a cost-push-created recession to run its course. How will that happen? Widespread layoffs, plant shutdowns, and business failures eventually occur. At some point the demand for oil, labor, and other inputs will decline so much that oil prices and nominal wages will decline. When that happens, the initial leftward shift of the short-run aggregate supply curve will reverse itself. That is, the declining per-unit production costs caused by the recession will shift the short-run aggregate supply curve rightward from AS_2 to AS_1. The price level will return to P_1, and the full-employment level of output will be restored at Q_f (point *a* on the long-run aggregate supply curve AS_{LR}).

This analysis yields two generalizations:

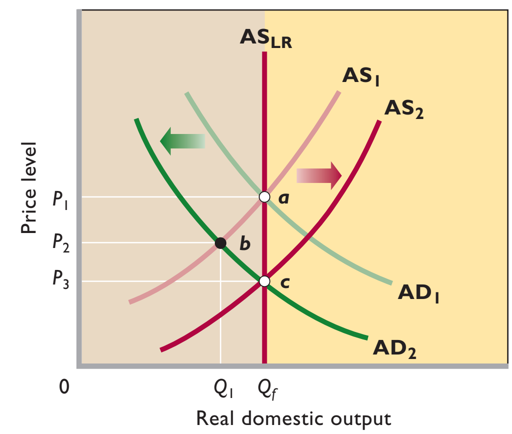
- If the government attempts to maintain full employment when there is cost-push inflation, even more inflation will occur.
- If the government takes a hands-off approach to cost-push inflation, the recession will linger. Although falling input prices will eventually undo the initial rise in per-unit production costs, the economy in the meantime will experience high unemployment and a loss of real output.

Recession and the Extended AD-AS Model

By far the most controversial application of the extended AD-AS model is its application to recession (or depression) caused by decreases in aggregate demand. We will look at this controversy in detail in Chapter 37; here we simply identify the key point of contention.

Suppose in Figure 36.5 that aggregate demand initially is AD_1 and that the short-run and long-run aggregate supply curves are AS_1 and AS_{LR}, respectively. Therefore, as shown by point *a*, the price level is P_1 and output is Q_f. Now suppose that investment spending declines dramatically, reducing aggregate demand to AD_2. Observe that real output declines from Q_f to Q_1, indicating that a recession has occurred. But if we make the controversial assumption that prices and wages are flexible downward, the price level falls from P_1 to P_2. With the economy producing below potential output at point *b*, demand for inputs will be low. Eventually, nominal wages themselves fall to restore the previous real wage; when that happens, the short-run aggregate supply curve shifts rightward from AS_1 to AS_2. The negative GDP gap evaporates without the need for expansionary fiscal or monetary policy since real output expands from Q_1 (point *b*) back to Q_f (point *c*). The economy is again located on its long-run aggregate supply curve AS_{LR}, but now at lower price level P_3.

FIGURE 36.5 Recession in the extended AD-AS model. A recession occurs when aggregate demand shifts leftward, as from AD_1 to AD_2. If prices and wages are downwardly flexible, the price level falls from P_1 to P_2 as the economy moves from point *a* to point *b*. With the economy in recession at point *b*, wages eventually fall, shifting the aggregate supply curve from AS_1 to AS_2. The price level declines to P_3, and real output returns to Q_f. The economy moves from *a* to *b* to *c*.

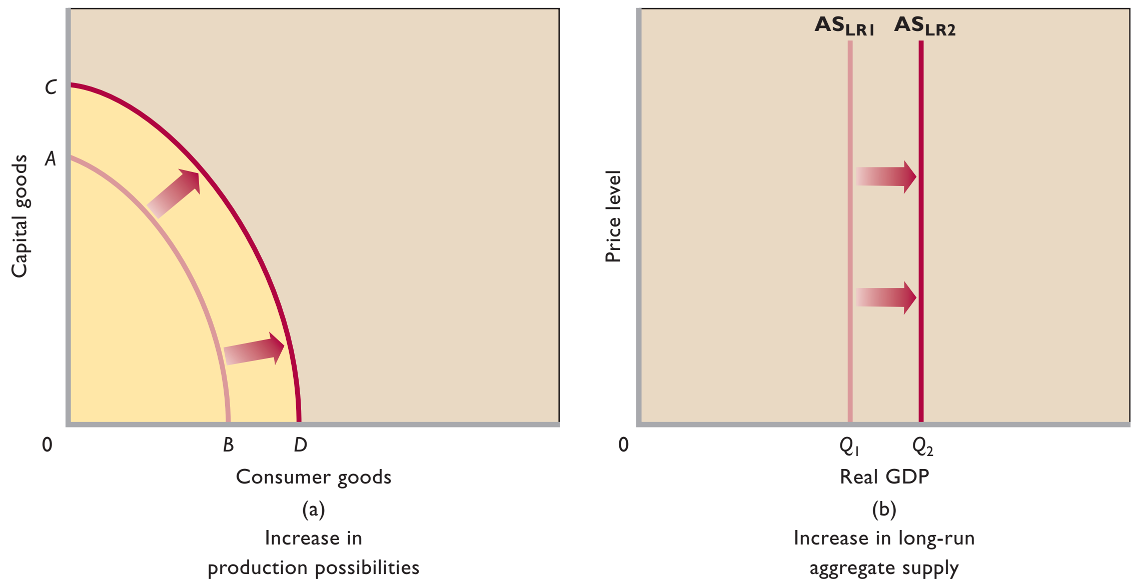


There is much disagreement about this hypothetical scenario. The key point of dispute revolves around the degree to which both input and output prices may be downwardly inflexible and how long it would take in the actual economy for the necessary downward price and wage adjustments to occur to regain the full-employment level of output. For now, suffice it to say that most economists believe that if such adjustments are forthcoming, they will occur only after the economy has experienced a relatively long-lasting recession with its accompanying high unemployment and large loss of output. The severity and length of the major recession of 2007–2009 has strengthened this view. Therefore, economists recommend active monetary policy, and perhaps fiscal policy, to counteract recessions.

Economic Growth with Ongoing Inflation

In our analysis so far, we have seen how demand and supply shocks can cause, respectively, demand-push inflation and cost-push inflation. But in these previous cases, the

FIGURE 36.6 Production possibilities and long-run aggregate supply. (a) Economic growth driven by supply factors (such as improved technologies or the use of more or better resources) shifts an economy's production possibilities curve outward, as from *AB* to *CD*. (b) The same factors shift the economy's long-run aggregate supply curve to the right, as from AS_{LR1} to AS_{LR2}.



extent of the inflation was *finite* because the size of the initial movement in either the AD curve or the AS curve was *limited*. For instance, in Figure 36.3, the aggregate demand curve shifts right by a limited amount, from AD_1 to AD_2. As the economy's equilibrium moves from *a* to *b* to *c*, the price level rises from P_1 to P_2 to P_3. During this transition, inflation obviously occurs because the price level is rising. But once the economy reaches its new equilibrium at point *c*, the price level remains constant at P_3 and no further inflation takes place. That is, the limited movement in aggregate demand causes a limited amount of inflation that ends when the economy returns to full employment.

But the modern economy almost always experiences continuous, but usually mild, positive rates of inflation. That can only happen with ongoing shifts in either the aggregate demand or long-run aggregate supply curves because any single, finite shift will only cause inflation of limited duration. This insight is crucial to understanding why modern economies usually experience ongoing inflation while achieving economic growth. Both aggregate demand and long-run aggregate supply increase over time in the actual economy, and inflation occurs because the increases in aggregate demand generally exceed the increases in long-run aggregate supply. It will be helpful to examine this point graphically.

Increases in Long-Run Aggregate Supply As discussed in Chapter 26, economic growth is driven by supply factors such as improved technologies and access to more or better resources. Economists illustrate economic growth as either an outward shift of an economy's production possibilities curve or as a rightward shift of its long-run aggregate supply curve. As shown in Figure 36.6, the outward shift of the production possibilities curve from *AB* to *CD* in graph *a* is equivalent to the rightward shift of the economy's long-run aggregate supply curve from AS_{LR1} to AS_{LR2} in graph *b*.

Let's simply transfer this rightward shift of the economy's long-run aggregate supply curve to Figure 36.7, which depicts economic growth in the United States in the context of the extended aggregate demand–aggregate supply model. Suppose the economy's long-run aggregate supply curve initially is AS_{LR1}, while its aggregate demand curve and short-run aggregate supply curve are AD_1 and AS_1, as shown. The equilibrium price level is P_1 and the equilibrium level of real output is Q_1.

Now let's assume that economic growth driven by changes in supply factors (quantity and quality of resources and technology) shifts the long-run aggregate supply curve rightward from AS_{LR1} to AS_{LR2} while the economy's aggregate demand curve remains at AD_1. Also, suppose that product and resource prices are flexible downward. The economy's potential output will expand, as reflected by the increase of available real output from Q_1 to Q_2. With aggregate demand constant at AD_1, the rightward shift of the long-run aggregate supply curve will lower the price level from P_1 to P_3. Taken alone,

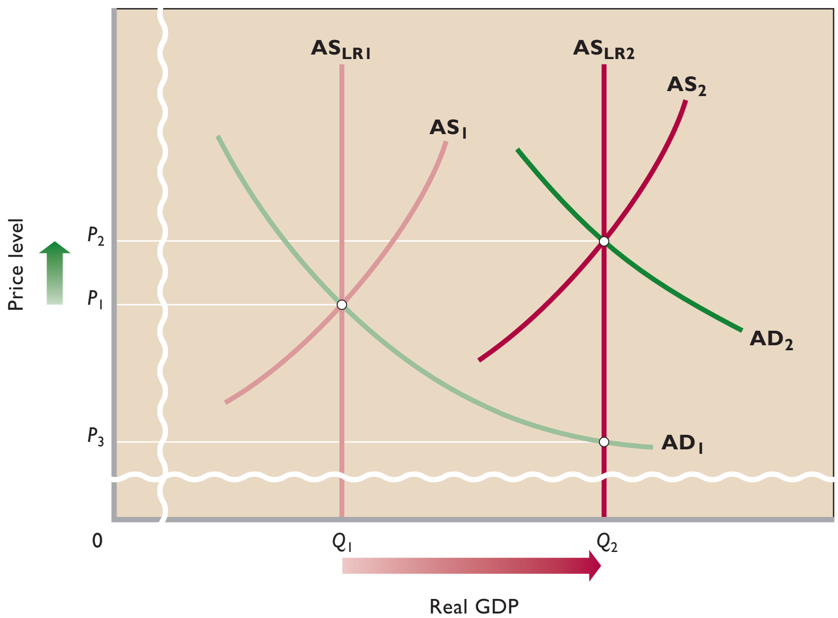


FIGURE 36.7 Depicting U.S. growth via the extended AD-AS model. Long-run aggregate supply and short-run aggregate supply have increased over time, as from AS_{LR1} to AS_{LR2} and AS_1 to AS_2. Simultaneously, aggregate demand has shifted rightward, as from AD_1 to AD_2. The actual outcome of these combined shifts has been economic growth, shown as the increase in real output from Q_1 to Q_2, accompanied by mild inflation, shown as the rise in the price level from P_1 to P_2.

expansions of long-run aggregate supply in the economy are deflationary.

Increases in Aggregate Demand and Inflation

But a decline in the price level, such as the one from P_1 to P_3 in Figure 36.7, is not part of the long-run U.S. growth experience. Why not? The answer is that the nation's central bank—the Federal Reserve—engineers ongoing increases in the nation's money supply to create rightward shifts of the aggregate demand curve. These increases in aggregate demand would be highly inflationary absent the increases in long-run aggregate supply. But because the Fed usually makes sure that the inflationary rightward shifts of the aggregate demand curve proceed only slightly faster than the deflationary rightward shifts of the aggregate supply curve, only mild inflation occurs along with economic growth.

We illustrate this outcome in Figure 36.7, where aggregate demand shifts to the right from AD_1 to AD_2 at the same time long-run aggregate supply shifts rightward from AS_{LR1} to AS_{LR2}. Real output expands from Q_1 to Q_2 and the price level increases from P_1 to P_2. At the higher price level P_2, the economy confronts a new short-run aggregate supply curve AS_2. The changes shown in Figure 36.7 describe the actual U.S. experience: economic growth, accompanied by mild inflation. Real output on average increases at about 3.4 percent annually and inflation averages 2 to 3 percent a year.

Of course, other long-term outcomes besides that depicted are entirely possible. Whether deflation, zero inflation, mild inflation, or rapid inflation accompanies economic growth depends on the extent to which aggregate demand increases relative to long-run aggregate supply. Over long periods, any inflation that accompanies economic growth is exclusively the result of aggregate demand increasing more rapidly than long-run aggregate supply. The expansion of long-run aggregate supply—of potential real GDP—is never the cause of inflation.

QUICK REVIEW 36.2

- In the short run, demand-pull inflation raises both the price level and real output; in the long run, nominal wages rise, the short-run aggregate supply curve shifts to the left, and only the price level increases.
- Cost-push inflation creates a policy dilemma for the government: If it engages in an expansionary policy to increase output, additional inflation will occur; if it does nothing, the recession will linger until input prices have fallen by enough to return the economy to producing at potential output.
- In the short run, a decline in aggregate demand reduces real output (creates a recession); in the long run, prices and nominal wages presumably fall, the short-run aggregate supply curve shifts to the right, and real output returns to its full-employment level.
- The economy has ongoing inflation because the Fed uses monetary policy to shift the AD curve to the right faster than the supply factors of economic growth shift the long-run AS curve to the right.

The Inflation-Unemployment Relationship

LO36.3 Explain the short-run trade-off between inflation and unemployment (the Phillips Curve).

We have just seen that the Fed can determine how much inflation occurs in the economy by how much it causes aggregate demand to shift relative to aggregate supply. Given that low inflation and low unemployment rates are the Fed's major economic goals, its ability to control inflation brings up at least two interesting policy questions: Are low unemployment and low inflation compatible goals or conflicting goals? What explains situations in which high unemployment and high inflation coexist?

The extended AD-AS model supports three significant generalizations relating to these questions:

- Under normal circumstances, there is a short-run trade-off between the rate of inflation and the rate of unemployment.
- Aggregate supply shocks can cause both higher rates of inflation and higher rates of unemployment.
- There is no significant trade-off between inflation and unemployment over long periods of time.

Let's examine each of these generalizations.

The Phillips Curve

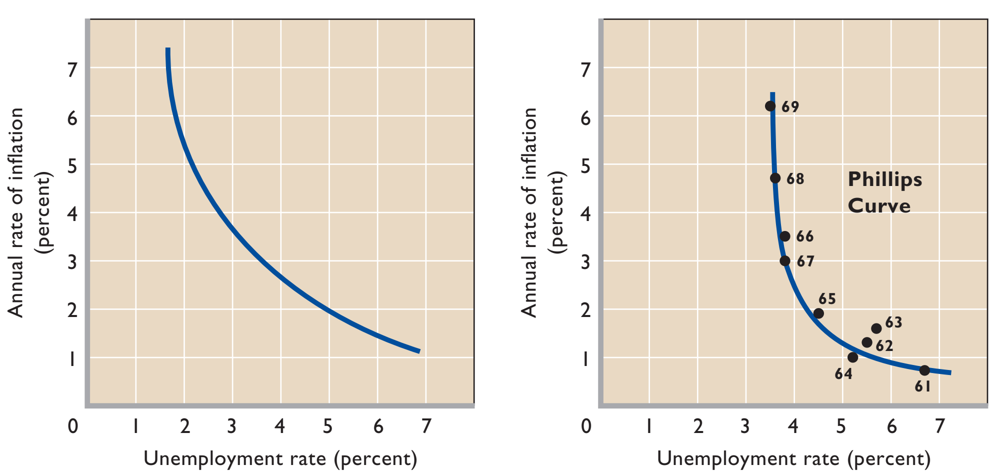
We can demonstrate the short-run trade-off between the rate of inflation and the rate of unemployment through the **Phillips Curve,** named after A. W. Phillips, who developed the idea in Great Britain. This curve, generalized in Figure 36.8a, suggests an inverse relationship between the rate of inflation and the rate of unemployment. Lower unemployment rates (measured as leftward movements on the horizontal axis) are associated with higher rates of inflation (measured as upward movements on the vertical axis).

ORIGIN OF THE IDEA

O36.1
Phillips Curve

The underlying rationale of the Phillips Curve becomes apparent when we view the short-run aggregate supply curve in Figure 36.9 and perform a simple mental experiment. Suppose that in some short-run period aggregate demand expands from AD_0 to AD_2, either because firms decide to buy more capital goods or the government decides to increase its expenditures. Whatever the cause, in the short run the price level rises from P_0 to P_2 and real output rises from Q_0 to Q_2. As real output rises, the unemployment rate falls.

FIGURE 36.8 The Phillips Curve: concept and empirical data. (a) The Phillips Curve relates annual rates of inflation and annual rates of unemployment for a series of years. Because this is an inverse relationship, there presumably is a trade-off between unemployment and inflation. (b) Data points for the 1960s seemed to confirm the Phillips Curve concept. (Note: The unemployment rates are annual averages and the inflation rates are on a December-to-December basis.)

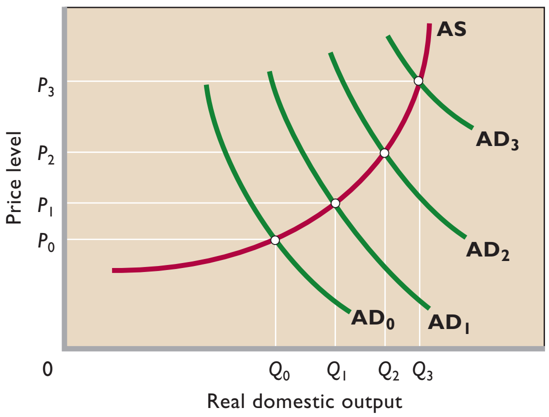


(a)
The concept

(b)
Data for the 1960s

Source: Bureau of Labor Statistics, **www.bls.gov**.

FIGURE 36.9 The short-run effect of changes in aggregate demand on real output and the price level. Comparing the effects of various possible increases in aggregate demand leads to the conclusion that the larger the increase in aggregate demand, the higher the rate of inflation and the greater the increase in real output. Because real output and the unemployment rate move in opposite directions, we can generalize that, given short-run aggregate supply, high rates of inflation should be accompanied by low rates of unemployment.



Now let's compare what would have happened if the increase in aggregate demand had been larger, say, from AD_0 to AD_3. The equilibrium at P_3 and Q_3 indicates that the amount of inflation and the growth of real output would both have been greater (and that the unemployment rate would have been lower). Similarly, suppose aggregate demand during the year had increased only modestly, from AD_0 to AD_1. Compared with our shift from AD_0 to AD_2, the amount of inflation and the growth of real output would have been smaller (and the unemployment rate higher).

The generalization we draw from this mental experiment is this: *Assuming a constant short-run aggregate supply curve*, high rates of inflation are accompanied by low rates of unemployment, and low rates of inflation are accompanied by high rates of unemployment. Other things equal, the expected relationship should look something like Figure 36.8a.

Figure 36.8b reveals that the facts for the 1960s nicely fit the theory. On the basis of that evidence and evidence from other countries, most economists working at the end of the 1960s concluded there was a stable, predictable trade-off between unemployment and inflation. Moreover, U.S. economic policy was built on that supposed trade-off. According to this thinking, it was impossible to achieve "full employment without inflation": Manipulation of aggregate demand through fiscal and monetary measures would simply move the economy along the Phillips Curve. An expansionary fiscal and monetary policy that boosted aggregate demand and lowered the unemployment rate would simultaneously increase inflation. A restrictive fiscal and monetary policy could be used to reduce the rate of inflation but only at the cost of a higher unemployment rate and more forgone production. Society had to choose between the incompatible goals of price stability and full employment; it had to decide where to locate on its Phillips Curve.

For reasons we will soon see, today's economists reject the idea of a stable, predictable Phillips Curve. Nevertheless, they agree there is a short-run trade-off between unemployment and inflation. *Given short-run aggregate supply*, increases in aggregate demand increase real output and reduce the unemployment rate. As the unemployment rate falls and dips below the natural rate, the excessive spending produces demand-pull inflation. Conversely, when recession sets in and the unemployment rate increases, the weak aggregate demand that caused the recession also leads to lower inflation rates.

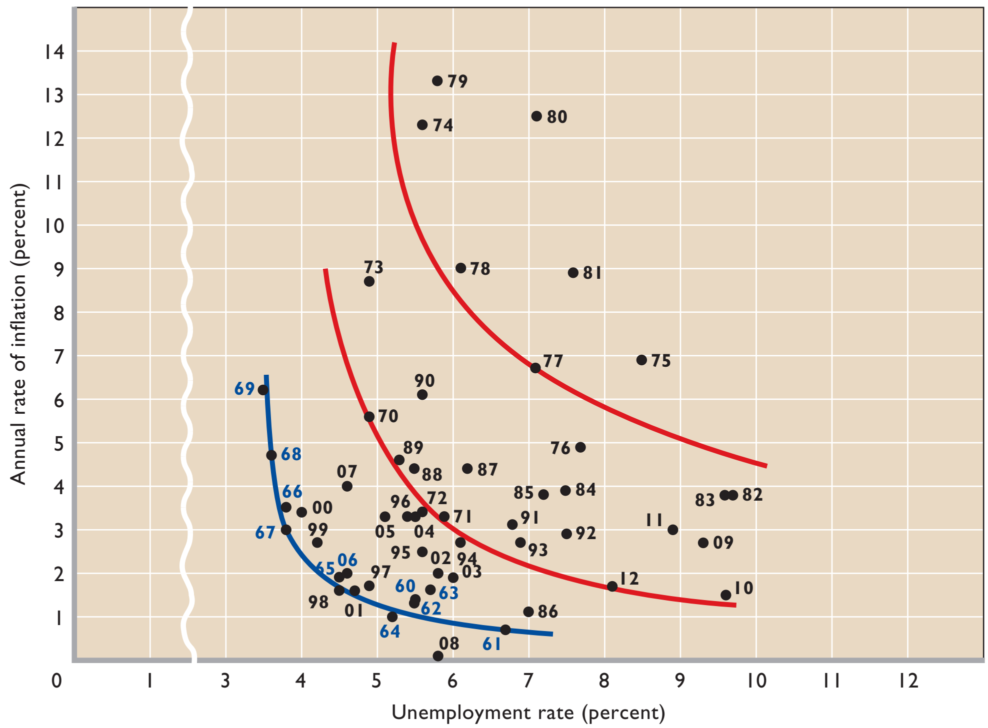
Periods of exceptionally low unemployment rates and inflation rates do occur, but only under special sets of economic circumstances. One such period was the late 1990s, when faster productivity growth increased aggregate supply and fully blunted the inflationary impact of rapidly rising aggregate demand (review Figure 30.11).

Aggregate Supply Shocks and the Phillips Curve

The unemployment-inflation experience of the 1970s and early 1980s demolished the idea of an always-stable Phillips Curve. In Figure 36.10 we show the Phillips Curve for the 1960s in blue and then add the data points for 1970 through 2012. Observe that in most of the years of the 1970s and early 1980s, the economy experienced both higher inflation rates and higher unemployment rates than it did in the 1960s. In fact, inflation and unemployment rose simultaneously in some of those years. This condition is called **stagflation**—a media term that combines the words "stagnation" and "inflation." If there still was any such thing as a Phillips Curve, it had clearly shifted outward, perhaps as shown.

Adverse Aggregate Supply Shocks The data points for the 1970s and early 1980s support our second generalization: Aggregate supply shocks can cause both higher rates of inflation and higher rates of unemployment. A series of adverse **aggregate supply shocks**—sudden, large increases in resource costs that jolt an economy's short-run aggregate supply curve leftward—hit the economy in the 1970s and early 1980s. The most significant of these shocks was a quadrupling of oil prices by the Organization of Petroleum Exporting Countries

FIGURE 36.10 Inflation rates and unemployment rates, 1960–2012. A series of aggregate supply shocks in the 1970s resulted in higher rates of inflation *and* higher rates of unemployment. So data points for the 1970s and 1980s tended to be above and to the right of the blue Phillips Curve for the 1960s. In the 1990s the inflation-unemployment data points slowly moved back toward the 1960s Phillips Curve. Points for the late 1990s and 2000s are similar to those from the 1960s. (Note: The unemployment rates are annual averages and the inflation rates are on a December-to-December basis.)



Source: Bureau of Labor Statistics, **www.bls.gov**.

(OPEC). Consequently, the cost of producing and distributing virtually every product and service rose rapidly. (Other factors working to increase U.S. costs during this period included major agricultural shortfalls, a greatly depreciated dollar, wage hikes previously held down by wage-price controls, and slower rates of productivity growth.)

These shocks shifted the aggregate supply curve to the left and distorted the usual inflation-unemployment relationship. Remember that we derived the inverse relationship between the rate of inflation and the unemployment rate shown in Figure 36.8a by shifting the aggregate demand curve along a stable short-run aggregate supply curve (Figure 36.9). But the cost-push inflation model shown in Figure 36.4 tells us that a *leftward shift* of the short-run aggregate supply curve increases the price level and reduces real output (and increases the unemployment rate). This, say most economists, is what happened in two periods in the 1970s. The U.S. unemployment rate shot up from 4.9 percent in 1973 to 8.5 percent in 1975, contributing to a significant decline in real GDP. In the same period, the U.S. price level rose by 21 percent. The stagflation scenario recurred in 1978, when OPEC increased oil prices by more than 100 percent. The U.S. price level rose by 26 percent over the 1978–1980 period, while unemployment increased from 6.1 to 7.1 percent.

Stagflation's Demise Another look at Figure 36.10 reveals a generally inward movement of the inflation-unemployment points between 1982 and 1989. By 1989 the lingering effects of the earlier period had subsided. One precursor to this favorable trend was the deep recession of 1981–1982, largely caused by a restrictive monetary policy aimed at reducing double-digit inflation. The recession upped the unemployment rate to 9.5 percent in 1982. With so many workers unemployed, those who were working accepted smaller increases in their nominal wages—or, in some cases, wage reductions—in order to preserve their jobs. Firms, in turn, restrained their price increases to try to retain their relative shares of a greatly diminished market.

Other factors were at work. Foreign competition throughout this period held down wage and price hikes in several basic industries such as automobiles and steel. Deregulation of the airline and trucking industries also resulted in wage reductions or so-called wage givebacks. A significant decline in OPEC's monopoly power and a greatly reduced reliance on oil in the production process produced a stunning fall in the price of oil and its derivative products, such as gasoline.

All these factors combined to reduce per-unit production costs and to shift the short-run aggregate supply curve rightward (as from AS_2 to AS_1 in Figure 36.4). Employment and output expanded, and the unemployment rate fell from 9.6 percent in 1983 to 5.3 percent in 1989. Figure 36.10 reveals that the inflation-unemployment points for recent years are closer to the points associated with the Phillips Curve of the 1960s than to the points in the late 1970s and early 1980s. Even the recession of 2007–2009 did not greatly alter the recent inflation rate–unemployment rate pattern. In 2008, the unemployment rate was 5.8 percent, but the inflation rate was near zero at just 0.1 percent. The inflation-unemployment point for 2009, however, moved up and to the right relative to the point for 2008. In 2009 unemployment was 9.3 percent, while the inflation rate on a December-to-December basis was 2.7 percent. As the economy slowly recovered after the Great Recession, the points for 2010, 2011, and 2012 gradually moved leftward as the unemployment rate slowly declined from 9.6 percent to 8.1 percent and the inflation rate remained moderate at between 1.5 percent and 3.0 percent.

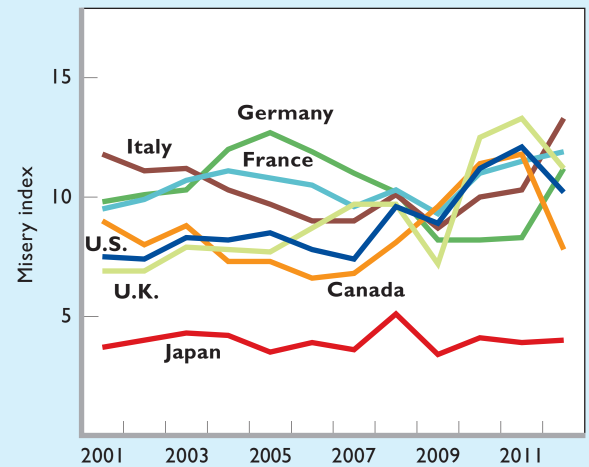
The media sometimes express the sum of the unemployment rate and the inflation rate as a rough gauge of the economic discomfort that inflation and unemployment jointly impose on an economy in a particular year. The sum of the two rates is used to compute the *misery index*. For example, a nation with a 5 percent unemployment rate and a 5 percent inflation rate has a misery index number of 10, as does a nation with an 8 percent unemployment rate and a 2 percent inflation rate.

Global Perspective 36.1 shows the misery index for several nations between 2001 and 2012. (It uses average annual inflation rates, not the December-to-December inflation rates used to construct Figure 36.10.) The U.S. misery index number has been neither exceptionally low nor exceptionally high relative to the misery index numbers for the other major economies shown. But bear in mind that economists do not put much stock in the misery index because they do not view the national discomfort associated with a 1 percent change in inflation and a 1 percent change in unemployment as necessarily equivalent. In particular, the misery index can greatly disguise the extent of hardship during a recession. The rise in the unemployment rate in these circumstances causes a huge loss of national output and income (and misery!), even if the higher unemployment rate is partially or fully offset with a decline in the rate of inflation.

GLOBAL PERSPECTIVE 36.1

The Misery Index, Selected Nations, 2001–2012

The U.S. misery index number (the sum of its unemployment rate and inflation rate) has generally been in the midrange of such numbers relative to other major economies in recent years.



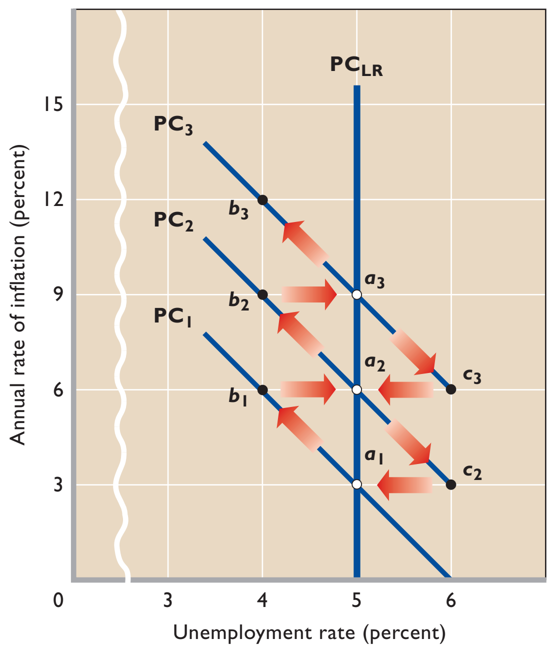
Source: Bureau of Labor Statistics, **stats.bls.gov**. Inflation rates and unemployment rates in this calculation are both on an average-annual basis.

The Long-Run Phillips Curve

LO36.4 Discuss why there is no long-run trade-off between inflation and unemployment.

The overall set of data points in Figure 36.10 supports our third generalization relating to the inflation-unemployment relationship: There is no apparent *long-run* trade-off between inflation and unemployment. Economists point out that when decades as opposed to a few years are considered, any rate of inflation is consistent with the natural rate of unemployment prevailing at that time. We know from Chapter 27 that the natural rate of unemployment is the unemployment rate that occurs when cyclical unemployment is zero; it is the full-employment rate of unemployment, or the rate of unemployment when the economy achieves its potential output.

FIGURE 36.11 The long-run vertical Phillips Curve. Increases in aggregate demand beyond those consistent with full-employment output may temporarily boost profits, output, and employment (as from a_1 to b_1). But nominal wages eventually will catch up so as to sustain real wages. When they do, profits will fall, negating the previous short-run stimulus to production and employment (the economy now moves from b_1 to a_2). Consequently, there is no trade-off between the rates of inflation and unemployment in the long run; that is, the long-run Phillips Curve is roughly a vertical line at the economy's natural rate of unemployment.



How can there be a short-run inflation-unemployment trade-off but not a long-run trade-off? Figure 36.11 provides the answer.

Short-Run Phillips Curve

Consider Phillips Curve PC_1 in Figure 36.11. Suppose the economy initially is experiencing a 3 percent rate of inflation and a 5 percent natural rate of unemployment. Such short-term curves as PC_1, PC_2, and PC_3 (drawn as straight lines for simplicity) exist because the actual rate of inflation is not always the same as the expected rate.

Establishing an additional point on Phillips Curve PC_1 will clarify this. We begin at a_1, where we assume nominal wages are set on the assumption that the 3 percent rate of inflation will continue. That is, because workers expect output prices to rise by 3 percent per year, they negotiate wage contracts that feature 3 percent per year increases in nominal wages so that these nominal wage increases will exactly offset the expected rise in prices and thereby keep their real wages the same.

But suppose that the rate of inflation rises to 6 percent, perhaps because the Fed has decided to move the AD curve to the right even faster than it had been before. With a nominal wage rate set on the expectation that the 3 percent rate of inflation will continue, the higher product prices raise business profits. Firms respond to the higher profits by hiring more workers and increasing output. In the short run, the economy moves to b_1, which, in contrast to a_1, involves a lower rate of unemployment (4 percent) and a higher rate of inflation (6 percent). The move from a_1 to b_1 is consistent both with an upsloping aggregate supply curve and with the inflation-unemployment trade-off implied by the Phillips Curve analysis. But this short-run Phillips Curve simply is a manifestation of the following principle: *When the actual rate of inflation is higher than expected, profits temporarily rise and the unemployment rate temporarily falls.*

Long-Run Vertical Phillips Curve

But point b_1 is not a stable equilibrium. Workers will recognize that their nominal wages have not increased as fast as inflation and will therefore renegotiate their labor contracts so that they feature faster increases in nominal wages. These faster increases in nominal wages make up for the higher rate of inflation and restore the workers' lost purchasing power. As these new labor contracts kick in, business profits will fall to their prior level. The reduction in profits means that the original motivation to employ more workers and increase output has disappeared.

Unemployment then returns to its natural level at point a_2. Note, however, that the economy now faces a higher actual and expected rate of inflation—6 percent rather than 3 percent. This happens because the new labor contracts feature 6 percent per year increases in wages to make up for the 6 percent per year inflation rate. Because wages are a production cost, this faster increase in wage rates will imply faster future increases in output prices as firms are forced to raise prices more rapidly to make up for the faster future rate of wage growth. Stated a bit differently, the initial increase in inflation will become *persistent* because it leads to renegotiated labor contracts that will perpetuate the higher rate of inflation. In addition, because the new labor contracts are public, it will also be the case that the higher rates of inflation that they will cause will be *expected* by everyone rather than being a surprise.

In view of the higher 6 percent expected rate of inflation, the short-run Phillips Curve shifts upward from PC_1 to PC_2 in Figure 36.11. An "along-the-Phillips-Curve" kind of move from a_1 to b_1 on PC_1 is merely a short-run or transient occurrence. In the long run, after nominal wage contracts catch up with increases in the inflation rate, unemployment returns to its natural rate at a_2, and there is a new short-run Phillips Curve PC_2 at the higher expected rate of inflation.

The scenario repeats if aggregate demand continues to increase. Prices rise momentarily ahead of nominal wages, profits expand, and employment and output increase (as implied by the move from a_2 to b_2). But, in time, nominal wages increase so as to restore real wages. Profits then fall to their original level, pushing employment back to the normal rate at a_3. The economy's "reward" for lowering the unemployment rate below the natural rate is a still higher (9 percent) rate of inflation.

Movements along the short-run Phillips curve (a_1 to b_1 on PC_1) cause the curve to shift to a less favorable position (PC_2, then PC_3, and so on). A stable Phillips Curve with the dependable series of unemployment-rate–inflation-rate trade-offs simply does not exist in the long run. The economy is characterized by a **long-run vertical Phillips Curve.**

The vertical line through a_1, a_2, and a_3 shows the long-run relationship between unemployment and inflation. Any rate of inflation is consistent with the 5 percent natural rate of unemployment. So, in this view, society ought to choose a low rate of inflation rather than a high one.

ORIGIN OF THE IDEA

O36.2
Long-run vertical Phillips Curve

Disinflation

The distinction between the short-run Phillips Curve and the long-run Phillips Curve also helps explain **disinflation**—reductions in the inflation rate from year to year. Suppose that in Figure 36.11 the economy is at a_3, where the inflation rate is 9 percent. And suppose that a decline in the rate at which aggregate demand shifts to the right faster than aggregate supply (as happened during the 1981–1982 recession) reduces inflation below the 9 percent expected rate, say, to 6 percent. Business profits fall because prices are rising less rapidly than wages. The nominal wage increases, remember, were set on the assumption that the 9 percent rate of inflation would continue. In response to the decline in profits, firms reduce their employment and consequently the unemployment rate rises. The economy temporarily slides downward from point a_3 to c_3 along the short-run Phillips Curve PC_3. *When the actual rate of inflation is lower than the expected rate, profits temporarily fall and the unemployment rate temporarily rises.*

Firms and workers eventually adjust their expectations to the new 6 percent rate of inflation, and thus newly negotiated wage increases decline. Profits are restored, employment rises, and the unemployment rate falls back to its natural rate of 5 percent at a_2. Because the expected rate of inflation is now 6 percent, the short-run Phillips Curve PC_3 shifts leftward to PC_2.

If the rate at which aggregate demand shifts to the right faster than aggregate supply declines even more, the scenario will continue. Inflation declines from 6 percent to, say, 3 percent, moving the economy from a_2 to c_2 along PC_2. The lower-than-expected rate of inflation (lower prices) squeezes profits and reduces employment. But, in the long run, firms respond to the lower profits by reducing their nominal wage increases. Profits are restored and unemployment returns to its natural rate at a_1 as the short-run Phillips Curve moves from PC_2 to PC_1. Once again, the long-run Phillips Curve is vertical at the 5 percent natural rate of unemployment.

QUICK REVIEW 36.3

- As implied by the upsloping short-run aggregate supply curve, there may be a short-run trade-off between the rate of inflation and the rate of unemployment. This trade-off is reflected in the Phillips Curve, which shows that lower rates of inflation are associated with higher rates of unemployment.
- Aggregate supply shocks that produce severe cost-push inflation can cause stagflation—simultaneous increases in the inflation rate and the unemployment rate. Such stagflation occurred from 1973–1975 and recurred from 1978–1980, producing Phillips Curve data points above and to the right of the Phillips Curve for the 1960s.
- After all nominal wage adjustments to increases and decreases in the rate of inflation have occurred, the economy ends up back at its full-employment level of output and its natural rate of unemployment. The long-run Phillips Curve therefore is vertical at the natural rate of unemployment.

Taxation and Aggregate Supply

LO36.5 Explain the relationship between tax rates, tax revenues, and aggregate supply.

A final topic in our discussion of aggregate supply is taxation, a key aspect of **supply-side economics.** "Supply-side economists" or "supply-siders" stress that changes in aggregate supply are an active force in determining the levels of inflation, unemployment, and economic growth. Government policies can either impede or promote rightward shifts of the short-run and long-run aggregate supply curves shown in Figure 36.2. One such policy is taxation.

These economists say that the enlargement of the U.S. tax system has impaired incentives to work, save, and invest. In this view, high tax rates impede productivity growth and hence slow the expansion of long-run aggregate supply. By reducing the after-tax rewards of workers and producers, high tax rates reduce the financial attractiveness of working, saving, and investing.

Supply-siders focus their attention on *marginal tax rates*—the rates on extra dollars of income—because those rates affect the benefits from working, saving, or investing more. In 2013 marginal federal income tax rates varied from 10 to 39.6 percent in the United States.

Taxes and Incentives to Work

Supply-siders believe that how long and how hard people work depends on the amounts of additional after-tax earnings they derive from their efforts. They say that lower marginal tax rates on earned incomes induce more work, and therefore increase aggregate inputs of labor. Lower marginal tax rates increase the after-tax wage rate and make leisure more expensive and work more attractive. The higher opportunity cost of leisure encourages people to substitute work for leisure. This increase in productive effort is achieved in many ways: by increasing the number of hours worked per day or week, by encouraging workers to postpone retirement, by inducing more people to enter the labor force, by motivating people to work harder, and by avoiding long periods of unemployment.

Incentives to Save and Invest

High marginal tax rates also reduce the rewards for saving and investing. For example, suppose that Tony saves \$10,000 at 8 percent interest, bringing him \$800 of interest per year. If his marginal tax rate is 40 percent, his after-tax interest earnings will be \$480, not \$800, and his after-tax interest rate will fall to 4.8 percent. While Tony might be willing to save (forgo current consumption) for an 8 percent return on his saving, he might rather consume when the return is only 4.8 percent.

Saving, remember, is the prerequisite of investment. Thus, supply-side economists recommend lower marginal tax rates on interest earned from saving. They also call for lower taxes on income from capital to ensure that there are ready investment outlets for the economy's enhanced pool of saving. A critical determinant of investment spending is the expected *after-tax* return on that spending.

To summarize: Lower marginal tax rates encourage saving and investing. Workers therefore find themselves equipped with more and technologically superior machinery and equipment. Labor productivity rises, and that expands long-run aggregate supply and economic growth, which in turn keeps unemployment rates and inflation low.

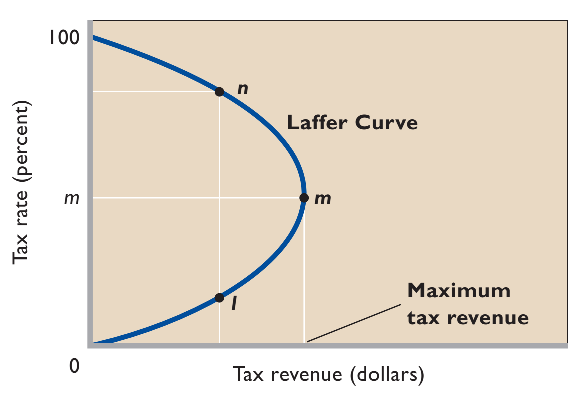
The Laffer Curve

In the supply-side view, reductions in marginal tax rates increase the nation's aggregate supply and can leave the nation's tax revenues unchanged or even enlarge them. Thus, supply-side tax cuts need not produce federal budget deficits.

This idea is based on the **Laffer Curve,** named after Arthur Laffer, who popularized it. As Figure 36.12 shows, the Laffer Curve depicts the relationship between tax rates and tax revenues. As tax rates increase from 0 to 100 percent, tax revenues increase from zero to some maximum level (at *m*) and then fall to zero. Tax revenues decline beyond some point because higher tax rates discourage economic activity, thereby shrinking the tax base (domestic output and income). This is easiest to see at the extreme, where the tax rate is 100 percent. Tax revenues here are, in theory, reduced to zero because the 100 percent confiscatory tax rate has halted production. A 100 percent tax rate applied to a tax base of zero yields no revenue.

In the early 1980s, Laffer suggested that the United States was at a point such as *n* on the curve in Figure 36.12.

FIGURE 36.12 The Laffer Curve. The Laffer Curve suggests that up to point *m* higher tax rates will result in larger tax revenues. But tax rates higher than *m* will adversely affect incentives to work and produce, reducing the size of the tax base (output and income) to the extent that tax revenues will decline. It follows that if tax rates are above *m*, reductions in tax rates will produce increases in tax revenues.



At *n*, tax rates are so high that production is discouraged to the extent that tax revenues are below the maximum at *m*. If the economy is at *n*, then lower tax rates can either increase tax revenues or leave them unchanged. For example, lowering the tax rate from point *n* to point *l* would bolster the economy such that the government would bring in the same total amount of tax revenue as before.

Laffer's reasoning was that lower tax rates stimulate incentives to work, save and invest, innovate, and accept business risks, thus triggering an expansion of real output and income. That enlarged tax base sustains tax revenues even though tax rates are lowered. Indeed, between *n* and *m* lower tax rates result in *increased* tax revenue.

Also, when taxes are lowered, tax avoidance (which is legal) and tax evasion (which is not) decline. High marginal tax rates prompt taxpayers to avoid taxes through various tax shelters, such as buying municipal bonds, on which the interest earned is tax-free. High rates also encourage some taxpayers to conceal income from the Internal Revenue Service. Lower tax rates reduce the inclination to engage in either tax avoidance or tax evasion.

The Laffer curve also implies that for any particular amount of tax revenue that the government can possibly collect, there will be both a high tax rate at which that amount of revenue can be collected as well as a low tax rate at which that amount of revenue can be collected. As an example, compare points *n* an *l* in Figure 36.12. Point *n* has a high tax rate and point *l* has a low tax rate, but they both collect the same amount of tax revenue. So if the government's major goal when setting tax rates is simply to collect a particular total amount of tax revenue, Laffer argued that the government should always opt for the lower tax rate. By doing so, the government would collect the revenue it desired while impinging as little as possible on the private economy.

CONSIDER THIS . . .

Sherwood Forest

The popularization of the idea that tax-rate reductions may increase tax revenues owes much to Arthur Laffer's ability to present his ideas simply. In explaining his thoughts to a *Wall Street Journal* editor over lunch, Laffer reportedly took out his pen and drew the curve on a napkin. The editor retained the napkin and later reproduced the curve in an editorial in *The Wall Street Journal*. The Laffer Curve was born. The idea it portrayed became the centerpiece of economic policy under the Reagan administration (1981–1989), which cut tax rates on personal income by 25 percent over a three-year period.

Laffer illustrated his supply-side views with a story relating to Robin Hood, who, you may recall, stole from the rich to give to the poor. Laffer likened people traveling through Sherwood Forest to taxpayers, whereas Robin Hood and his band of merry men were government. As taxpayers passed through the forest, Robin Hood and his men intercepted them and forced them to hand over their money. Laffer asked audiences, "Do you think that travelers continued to go through Sherwood Forest?"

The answer he sought and got, of course, was "no." Taxpayers will avoid Sherwood Forest to the greatest extent possible. They will lower their taxable income by reducing work hours, retiring earlier, saving less, and engaging in tax avoidance and tax evasion activities. Robin Hood and his men may end up with less revenue than if they collected a relatively small "tax" from each traveler for passage through the forest.

Criticisms of the Laffer Curve

The Laffer Curve and its supply-side implications have been subject to severe criticism.

Taxes, Incentives, and Time A fundamental criticism relates to the degree to which economic incentives are sensitive to changes in tax rates. Skeptics say ample empirical evidence shows that the impact of a tax cut on incentives is small, of uncertain direction, and relatively slow to emerge. For example, with respect to work incentives, studies indicate that decreases in tax rates lead some people to work more but lead others to work less. Those who work more are enticed by the higher after-tax pay; they substitute work for leisure because the opportunity cost of leisure has increased. But other people work less because the higher after-tax pay enables them to "buy more leisure." With the tax cut, they can earn the same level of after-tax income as before with fewer work hours.

Inflation or Higher Real Interest Rates Most economists think that the demand-side effects of a tax cut are more immediate and certain than longer-term supply-side effects. Thus, tax cuts undertaken when the economy is at or near full employment may produce increases in aggregate demand that overwhelm any increase in aggregate supply. The likely result is inflation or restrictive monetary policy to prevent it. If the latter, real interest rates will rise and investment will decline. This will defeat the purpose of the supply-side tax cuts.

Position on the Curve Skeptics say that the Laffer Curve is merely a logical proposition and assert that there must be some level of tax rates between 0 and 100 percent at which tax revenues will be at their maximum. Economists of all persuasions can agree with this. But the issue of where a particular economy is located on its Laffer Curve is an empirical question. If we assume that we are at point *n* in Figure 36.12, then tax cuts will increase tax revenues. But if the economy is at any point below *m* on the curve, tax-rate reductions will reduce tax revenues.

Rebuttal and Evaluation

Supply-side advocates respond to the skeptics by contending that the Reagan tax cuts in the 1980s worked as Laffer predicted. Although the top marginal income tax rates on earned income were cut from 50 to 28 percent in that decade, real GDP and tax revenues were substantially higher at the end of the 1990s than at the beginning.

But the general view among economists is that the Reagan tax cuts, coming at a time of severe recession, helped boost aggregate demand and return real GDP to its full-employment output and normal growth path. As the economy expanded, so did tax revenues despite the lower tax rates. The rise in tax revenues caused by economic growth swamped the declines in revenues from lower tax rates. In essence, the Laffer Curve shown in Figure 36.12 stretched rightward, increasing net tax revenues. But the tax-rate cuts did not produce extraordinary rightward shifts of the long-run aggregate supply curve. Indeed, saving fell as a percentage of personal income during the period and productivity growth was sluggish. Real GDP sprang back vigorously from recessionary levels, but the economic growth rate soon reverted back to its longer-term average.

Because government expenditures rose more rapidly than tax revenues in the 1980s, large budget deficits occurred. In 1993 the Clinton administration increased the top marginal tax rates from 31 to 39.6 percent to address these deficits. The economy boomed in the last half of the 1990s, and by the end of the decade tax revenues were so high relative to government expenditures that budget surpluses emerged. In 2001 the Bush administration reduced marginal tax rates over a series of years, partially "to return excess revenues to taxpayers." In 2003 the top marginal tax rate fell to 35 percent. Also, the income tax rates on capital gains and dividends were reduced to 15 percent. Economists generally agree that the Bush tax cuts, along with a highly expansionary monetary policy, helped revive and expand the economy following the recession of 2001. Strong growth of output and income in 2004 and 2005 produced large increases in tax revenues, although large budget deficits remained because spending also increased rapidly. The 2004 deficit was $413 billion and the 2005 deficit was $318 billion. The deficit fell over the next two years, to $162 billion in 2007. But the previously discussed financial crisis plunged the economy into a severe recession beginning in December 2007 and lasting into 2009. That caused income to fall and tax revenues to plummet. Also, the government implemented highly expansionary fiscal policy. Budget deficits increased to $459 billion in 2008 and $1.4 trillion in 2009.

Today, there is general agreement that the U.S. economy is operating at a point below *m*—rather than above *m*—on the Laffer Curve in Figure 36.12. In this zone, the overall effect is that personal tax-rate increases raise tax revenues while personal tax-rate decreases reduce tax revenues. But at the same time, economists recognize that, other things being equal, cuts in tax rates reduce tax revenues in percentage terms by less than the tax-rate reductions. Similarly, tax-rate increases do not raise tax revenues by as much in percentage terms as the tax-rate increases. This is true because changes in marginal tax rates alter taxpayer behavior and thus affect taxable income. Although these effects seem to be relatively modest, they need to be considered in designing tax policy—and, in fact, the federal government's Office of Tax Policy created a special division in 2007 devoted to estimating the magnitude of such effects when it comes to proposed changes in U.S. tax laws. Thus, supply-side economics has contributed to how economists and policymakers design and implement fiscal policy.

QUICK REVIEW 36.4

- Supply-side economists focus their attention on government policies, such as high taxation, that may impede the expansion of aggregate supply.
- The Laffer Curve relates tax rates to levels of tax revenue and suggests that, under some circumstances, cuts in tax rates will expand the tax base (output and income) and increase tax revenues.
- Most economists believe that the United States is currently operating in the range of the Laffer Curve where tax rates and tax revenues move in the same, not opposite, directions.
- Today's economists recognize the importance of considering supply-side effects in designing optimal fiscal policy.

LAST WORD

Do Tax Increases Reduce Real GDP?*

Determining the Relationship Between Changes in Taxes and Permanent Changes in Real GDP is Fraught with Complexities and Difficulties. University of California-Berkeley Economists Christina Romer and David Romer have Recently Devised a Novel Way to Approach the Topic. Their Findings Suggest That Tax Increases Reduce Real GDP.†

How do changes in the level of taxation affect the level of economic activity? The simple correlation between taxation and economic activity shows that, on average, when economic activity rises more rapidly, tax revenues also are rising more rapidly. But this correlation almost surely does not reflect a positive effect of tax increases on output. Rather, under our tax system, any positive shock to output raises tax revenues by increasing income.

In "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks," authors Christina Romer and David Romer observe that this difficulty is just one of many manifestations of a more general problem. Changes in taxes occur for many reasons. And, because the factors that give rise to tax changes often are correlated with other developments in the economy, disentangling the effects of the tax changes from the other effects of these underlying factors is inherently difficult.

To address this problem, Romer and Romer use the narrative record—presidential speeches, executive branch documents, congressional reports, and so on—to identify the size, timing, and principal motivation for all major tax policy actions in the post–World War II United States. This narrative analysis allows them to separate revenue changes resulting from legislation from changes occurring for other reasons. It also allows them to classify legislated changes according to their primary motivation.

Romer and Romer find that despite the complexity of the legislative process, most significant tax changes have been motivated by one of four factors: counteracting other influences in the economy; paying for increases in government spending (or lowering taxes in conjunction with reductions in spending); addressing an inherited budget deficit; and promoting long-run growth. They observe that legislated tax changes taken to counteract other influences on the economy, or to pay for increases in government spending, are very likely to be correlated with other factors affecting the economy. As a result these observations are likely to lead to unreliable estimates of the effect of tax changes.

Tax changes that are made to promote long-run growth, or to reduce an inherited budget deficit, in contrast, are undertaken for reasons essentially unrelated to other factors influencing output. Thus, examining the behavior of output following these tax changes is likely to provide more reliable estimates of the output effects of tax changes. *The results of this more reliable test indicate that tax changes have very large effects: a tax increase of 1 percent of GDP lowers real GDP by roughly 2 to 3 percent.*

These output effects are highly persistent. The behavior of inflation and unemployment suggests that this persistence reflects long-lasting departures of output from previous levels. Romer and Romer also find that output effects of tax changes are much more closely tied to the actual changes in taxes than news about future changes, and that investment falls sharply in response to tax changes. Indeed, the strong response of investment helps to explain why the output consequences of tax increases are so large.

Romer and Romer find suggestive evidence that tax increases to reduce an inherited budget deficit have much smaller output costs than other tax increases. This is consistent with the idea that deficit-driven tax increases may have important expansionary effects through [improved] expectations and [lower] long-term interest rates, or through [enhanced] confidence.

---

*Abridged from Les Picker, "Tax Increases Reduce GDP," *The NBER Digest*, February/March 2008. The *Digest* provides synopses of research papers in progress by economists affiliated with the National Bureau of Economic Research (NBER).

†Christina Romer and David Romer, "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks," *American Economic Review*, June 2010, pp. 763–801.

SUMMARY

LO36.1 Explain the relationship between short-run aggregate supply and long-run aggregate supply.

In macroeconomics, the short run is a period in which nominal wages do not respond to changes in the price level. In contrast, the long run is a period in which nominal wages fully respond to changes in the price level.

The short-run aggregate supply curve is upsloping. Because nominal wages are unresponsive to price-level changes, increases in the price level (prices received by firms) increase profits and real output. Conversely, decreases in the price level reduce profits and real output. However, the long-run aggregate supply curve is vertical. With sufficient time for adjustment, nominal wages rise and fall with the price level, moving the economy along a vertical aggregate supply curve at the economy's full-employment output.

LO36.2 Discuss how to apply the "extended" (short-run/long-run) AD-AS model to inflation, recessions, and economic growth.

In the short run, demand-pull inflation raises the price level and real output. Once nominal wages rise to match the increase in the price level, the temporary increase in real output is reversed.

In the short run, cost-push inflation raises the price level and lowers real output. Unless the government expands aggregate demand, nominal wages will eventually decline under conditions of recession and the short-run aggregate supply curve will shift back to its initial location. Prices and real output will eventually return to their original levels.

If prices and wages are flexible downward, a decline in aggregate demand will lower output and the price level. The decline in the price level will eventually lower nominal wages and shift the short-run aggregate supply curve rightward. Full-employment output will thus be restored.

One-time changes in aggregate demand (AD) and aggregate supply (AS) can only cause limited bouts of inflation. Ongoing mild inflation occurs because the Fed purposely increases AD slightly faster than the expansion of long-run AS (driven by the supply factors of economic growth).

LO36.3 Explain the short-run trade-off between inflation and unemployment (the Phillips Curve).

Assuming a stable, upsloping short-run aggregate supply curve, rightward shifts of the aggregate demand curve of various sizes yield the generalization that high rates of inflation are associated with low rates of unemployment, and vice versa. This inverse relationship is known as the Phillips Curve, and empirical data for the 1960s seemed to be consistent with it.

In the 1970s and early 1980s the Phillips Curve apparently shifted rightward, reflecting stagflation—simultaneously rising inflation rates and unemployment rates. The higher unemployment rates and inflation rates resulted mainly from huge oil price increases that caused large leftward shifts in the short-run aggregate supply curve (so-called aggregate supply shocks). The Phillips Curve shifted inward toward its original position in the 1980s. By 1989 stagflation had subsided, and the data points for the late 1990s and first half of the first decade of the 2000s were similar to those of the 1960s. The new pattern continued until 2009, when the unemployment rate jumped to 9.3 percent and the inflation rate rose on a December-to-December basis to 2.7 percent.

LO36.4 Discuss why there is no long-run trade-off between inflation and unemployment.

Although there is a short-run trade-off between inflation and unemployment, there is no long-run trade-off. Workers will adapt their expectations to new inflation realities, and when they do, the unemployment rate will return to the natural rate. So the long-run Phillips Curve is vertical at the natural rate, meaning that higher rates of inflation do not permanently "buy" the economy less unemployment.

LO36.5 Explain the relationship between tax rates, tax revenues, and aggregate supply.

Supply-side economists focus attention on government policies, such as high taxation, that impede the expansion of aggregate supply. The Laffer Curve relates tax rates to levels of tax revenue and suggests that under some circumstances, cuts in tax rates will expand the tax base (output and income) and increase tax revenues. Most economists, however, believe that the United States is currently operating in the range of the Laffer Curve where tax rates and tax revenues move in the same, not opposite, direction.

Today's economists recognize the importance of considering supply-side effects in designing optimal fiscal policy.

TERMS AND CONCEPTS

short run
long run
Phillips Curve
stagflation
aggregate supply shocks
long-run vertical Phillips Curve
disinflation
supply-side economics
Laffer Curve

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Distinguish between the short run and the long run as they relate to macroeconomics. Why is the distinction important? LO36.1
2. Which of the following statements are true? Which are false? Explain why the false statements are untrue. LO36.1
   a. Short-run aggregate supply curves reflect an inverse relationship between the price level and the level of real output.
   b. The long-run aggregate supply curve assumes that nominal wages are fixed.
   c. In the long run, an increase in the price level will result in an increase in nominal wages.
3. Suppose the government misjudges the natural rate of unemployment to be much lower than it actually is, and thus undertakes expansionary fiscal and monetary policies to try to achieve the lower rate. Use the concept of the short-run Phillips Curve to explain why these policies might at first succeed. Use the concept of the long-run Phillips Curve to explain the long-run outcome of these policies. LO36.4
4. What do the distinctions between short-run aggregate supply and long-run aggregate supply have in common with the distinction between the short-run Phillips Curve and the long-run Phillips Curve? Explain. LO36.4
5. What is the Laffer Curve, and how does it relate to supply-side economics? Why is determining the economy's location on the curve so important in assessing tax policy? LO36.5
6. Why might one person work more, earn more, and pay more income tax when his or her tax rate is cut, while another person will work less, earn less, and pay less income tax under the same circumstance? LO36.5
7. LAST WORD On average, does an increase in taxes raise or lower real GDP? If taxes as a percent of GDP go up 1 percent, by how much does real GDP change? Are the decreases in real GDP caused by tax increases temporary or permanent? Does the intention of a tax increase matter?

REVIEW QUESTIONS

1. Suppose the full-employment level of real output (*Q*) for a hypothetical economy is $250 and the price level (*P*) initially is 100. Use the short-run aggregate supply schedules below to answer the questions that follow: LO36.1

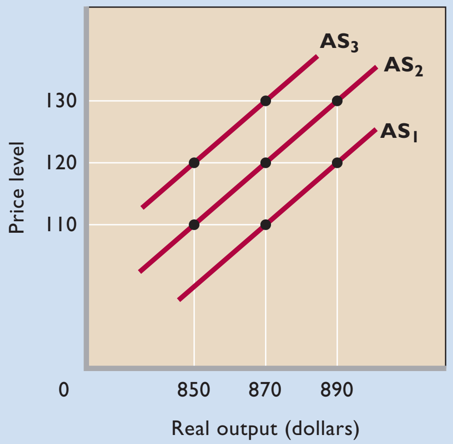
AS (P_{100})		AS (P_{125})		AS (P_{75})	
P	*Q*	*P*	*Q*	*P*	*Q*
125	$280	125	$250	125	$310
100	250	100	220	100	280
75	220	75	190	75	250

   a. What will be the level of real output in the short run if the price level unexpectedly rises from 100 to 125 because of an increase in aggregate demand? What if the price level unexpectedly falls from 100 to 75 because of a decrease in aggregate demand? Explain each situation, using numbers from the table.
   b. What will be the level of real output in the long run when the price level rises from 100 to 125? When it falls from 100 to 75? Explain each situation.
   c. Show the circumstances described in parts *a* and *b* on graph paper, and derive the long-run aggregate supply curve.
2. Suppose that AD and AS intersect at an output level that is higher than the full-employment output level. After the economy adjusts back to equilibrium in the long run, the price level will be ____________. LO36.2
   a. Higher than it is now.
   b. Lower than it is now.
   c. The same as it is now.
3. Suppose that an economy begins in long-run equilibrium before the price level and real GDP both decline simultaneously. If those changes were caused by only one curve shifting, then those changes are best explained as the result of: LO36.2
   a. The AD curve shifting right.
   b. The AS curve shifting right.
   c. The AD curve shifting left.
   d. The AS curve shifting left.
4. Identify the two descriptions below as being the result of either cost-push inflation or demand-pull inflation. LO36.2
   a. Real GDP is below the full-employment level and prices have risen recently.
   b. Real GDP is above the full-employment level and prices have risen recently.
5. Use graphical analysis to show how each of the following would affect the economy first in the short run and then in the long run. Assume that the United States is initially operating at its full-employment level of output, that prices and wages are eventually flexible both upward and downward, and that there is no counteracting fiscal or monetary policy. LO36.2
   a. Because of a war abroad, the oil supply to the United States is disrupted, sending oil prices rocketing upward.
   b. Construction spending on new homes rises dramatically, greatly increasing total U.S. investment spending.
   c. Economic recession occurs abroad, significantly reducing foreign purchases of U.S. exports.

6. Between 1990 and 2009, the U.S. price level rose by about 64 percent while real output increased by about 62 percent. Use the aggregate demand–aggregate supply model to illustrate these outcomes graphically. **LO36.2**
7. Assume there is a particular short-run aggregate supply curve for an economy and the curve is relevant for several years. Use the AD-AS analysis to show graphically why higher rates of inflation over this period would be associated with lower rates of unemployment, and vice versa. What is this inverse relationship called? **LO36.3**
8. Aggregate supply shocks can cause ________ rates of inflation that are accompanied by ________ rates of unemployment. **LO36.3**
   a. Higher; higher.
   b. Higher; lower.
   c. Lower; higher.
   d. Lower; lower.
9. Suppose that firms are expecting 6 percent inflation while workers are expecting 9 percent inflation. How much of a pay raise will workers demand if their goal is to maintain the purchasing power of their incomes? **LO36.4**
   a. 3 percent.
   b. 6 percent.
   c. 9 percent.
   d. 12 percent.
10. Suppose that firms were expecting inflation to be 3 percent, but then it actually turned out to be 7 percent. Other things equal, firm profits will be: **LO36.4**
   a. Smaller than expected.
   b. Larger than expected.

PROBLEMS

1. Use the figure below to answer the following questions. Assume that the economy initially is operating at price level 120 and real output level $870. This output level is the economy's potential (or full-employment) level of output. Next, suppose that the price level rises from 120 to 130. By how much will real output increase in the short run? In the long run? Instead, now assume that the price level dropped from 120 to 110. Assuming flexible product and resource prices, by how much will real output fall in the short run? In the long run? What is the long-run level of output at each of the three price levels shown? **LO36.1**



2. **ADVANCED ANALYSIS** Suppose that the equation for a particular short-run AS curve is $P = 20 + 0.5Q$, where P is the price level and Q is real output in dollar terms. What is Q if the price level is 120? Suppose that the Q in your answer is the full-employment level of output. By how much will Q increase *in the short run* if the price level unexpectedly rises from 120 to 132? By how much will Q increase *in the long run* due to the price level increase? **LO36.1**
3. Suppose that over a 30-year period Buskerville's price level increased from 72 to 138, while its real GDP rose from $1.2 trillion to $2.1 trillion. Did economic growth occur in Buskerville? If so, by what average yearly rate in percentage terms (rounded to one decimal place)? Did Buskerville experience inflation? If so, by what average yearly rate in percentage terms (rounded to one decimal place)? Which shifted rightward faster in Buskerville: its long-run aggregate supply curve (AS_{LR}) or its aggregate demand curve (AD)? **LO36.2**
4. Suppose that for years East Confetti's short-run Phillips Curve was such that each 1 percentage point increase in its unemployment rate was associated with a 2 percentage point decline in its inflation rate. Then, during several recent years, the short-run pattern changed such that its inflation rate rose by 3 percentage points for every 1 percentage point drop in its unemployment rate. Graphically, did East Confetti's Phillips Curve shift upward or did it shift downward? **LO36.3**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **37**

Current Issues in Macro Theory and Policy

Learning Objectives

LO37.1 Describe alternative perspectives on the causes of macroeconomic instability, including the views of mainstream economists, monetarists, real-business-cycle advocates, and proponents of coordination failures.

LO37.2 Discuss why new classical economists believe the economy will "self-correct" from aggregate demand and aggregate supply shocks.

LO37.3 Identify and describe the variations of the debate over "rules" versus "discretion" in conducting stabilization policy.

LO37.4 Summarize the fundamental ideas and policy implications of mainstream macroeconomics, monetarism, and rational expectations theory.

As any academic discipline evolves, it naturally evokes a number of internal disagreements. Economics is no exception. In this chapter we examine a few alternative perspectives on macro theory and policy. We focus on the disagreements that various economists have about the answers to three interrelated questions: (1) What causes instability in the economy? (2) Is the economy self-correcting? (3) Should government adhere to *rules* or use *discretion* in setting economic policy?

What Causes Macro Instability?

LO37.1 Describe alternative perspectives on the causes of macroeconomic instability, including the views of mainstream economists, monetarists, real-business-cycle advocates, and proponents of coordination failures.

As earlier chapters have indicated, capitalist economies experienced considerable instability during the twentieth century. The United States, for example, experienced the Great Depression, numerous recessions, and periods of inflation. This instability greatly moderated between the early 1980s and 2007, but then the deep recession of 2007–2009 occurred. Economists have different perspectives about why instability like this happens.

Mainstream View

For simplicity, we will use the term "mainstream view" to characterize the prevailing macroeconomic perspective of the majority of economists. According to that view, instability in the economy arises from two sources: (1) price stickiness and (2) unexpected shocks to either aggregate demand or aggregate supply.

As we explained in detail in Chapter 36, in the long run, when both input and output prices are fully flexible and have time to adjust to any changes in aggregate demand or short-run aggregate supply, the economy will always return to producing at potential output. In the shorter run, however, stickiness in either input or output prices will mean that any shock to either aggregate demand or aggregate supply will result in changes in output and employment. Although they are not new to you, let's quickly review shocks to aggregate demand and aggregate supply.

Changes in Aggregate Demand Mainstream macroeconomics focuses on aggregate spending and its components. Recall that the basic equation underlying aggregate expenditures is

$$C_a + I_g + X_n + G = \text{GDP}$$

That is, the aggregate amount of after-tax consumption, gross investment, net exports, and government spending determines the total amount of goods and services produced and sold. In equilibrium, $C_a + I_g + X_n + G$ (aggregate expenditures) is equal to GDP (real output). A decrease in the price level increases equilibrium GDP and thus allows us to trace out a downsloping aggregate demand curve for the economy (see the appendix to Chapter 30). Any change in one of the spending components in the aggregate expenditures equation shifts the aggregate demand curve. This, in turn, changes equilibrium real output, the price level, or both.

Investment spending in particular is subject to wide "booms" and "busts." Significant increases in investment spending are multiplied into even greater increases in aggregate demand and thus can produce demand-pull inflation. In contrast, significant declines in investment spending are multiplied into even greater decreases in aggregate demand and thus can cause recessions.

Adverse Aggregate Supply Shocks In the mainstream view, the second source of macroeconomic instability arises on the supply side. Occasionally, such external events as wars or an artificial supply restriction of a key resource can boost resource prices and significantly raise per-unit production costs. The result is a sizable decline in a nation's aggregate supply, which destabilizes the economy by simultaneously causing cost-push inflation and recession.

Monetarist View

Monetarism (1) focuses on the money supply, (2) holds that markets are highly competitive, and (3) says that a competitive market system gives the economy a high degree of macroeconomic stability. Monetarists argue that the price and wage flexibility provided by competitive markets should cause fluctuations in aggregate demand to alter product and resource prices rather than output and employment. Thus, the market system would provide substantial macroeconomic stability *were it not for government interference in the economy.*

ORIGIN OF THE IDEA

O37.1
Monetarism

The problem, as monetarists see it, is that government has promoted downward wage inflexibility through the minimum-wage law, pro-union legislation, guaranteed prices for certain farm products, pro-business monopoly legislation, and so forth. The free-market system is capable of providing macroeconomic stability, but, despite good intentions, government interference has undermined that capability. Moreover, monetarists say that government has contributed to the economy's business cycles through its clumsy and mistaken attempts to achieve greater stability through its monetary policies.

Equation of Exchange The fundamental equation of monetarism is the **equation of exchange:**

$$MV = PQ$$

where M is the supply of money; V is the **velocity** of money, that is, the average number of times per year a dollar is

ORIGIN OF THE IDEA

O37.2
Equation of exchange

spent on final goods and services; P is the price level or, more specifically, the average price at which each unit of physical output is sold; and Q is the physical volume of all goods and services produced.

The left side of the equation of exchange, MV, represents the total amount spent by purchasers of output, while the right side, PQ, represents the total amount received by sellers of that output. The nation's money supply (M) multiplied by the number of times it is spent each year (V) must equal the nation's nominal GDP ($= P \times Q$). The dollar value of total spending has to equal the dollar value of total output.

Stable Velocity Monetarists say that velocity, V, in the equation of exchange is relatively stable. To them, "stable" is not synonymous with "constant," however. Monetarists are aware that velocity has generally trended upward over the last several decades. Shorter pay periods, widespread use of credit cards, and faster means of making payments enable people to hold less money and to turn it over more rapidly than was possible in earlier times. These factors have enabled people to reduce their holdings of cash and checkbook money relative to the size of the nation's nominal GDP.

When monetarists say that velocity is stable, they mean that the factors altering velocity change gradually and predictably and that changes in velocity from one year to the next can be readily anticipated. Moreover, they hold that velocity does not change in response to changes in the money supply itself. Instead, people have a stable desire to hold money relative to holding other financial assets, holding real assets, and buying current output. The factors that determine the amount of money the public wants to hold depend mainly on the level of nominal GDP.

Example: Assume that when the level of nominal GDP is \$400 billion, the public desires \$100 billion of money to purchase that output. That means that V is 4 (= \$400 billion of nominal GDP/\$100 billion of money). If we further assume that the actual supply of money is \$100 billion, the economy is in equilibrium with respect to money; the actual amount of money supplied equals the amount the public wants to hold.

If velocity is stable, the equation of exchange suggests that there is a predictable relationship between the money supply and nominal GDP ($= PQ$). An increase in the money supply of, say, \$10 billion would upset equilibrium in our example since the public would find itself holding more money or liquidity than it wants. That is, the actual amount of money held (\$110 billion) would exceed the amount of holdings desired (\$100 billion). In that case, the reaction of the public (households and businesses) is to restore its desired balance of money relative to other items, such as stocks and bonds, factories and equipment, houses and automobiles, and clothing and toys. But the spending of money by individual households and businesses would leave more cash in the checkable deposits or billfolds of other households and firms. And they too would try to "spend down" their excess cash balances. But, overall, the \$110 billion supply of money cannot be spent down because a dollar spent is a dollar received.

Instead, the collective attempt to reduce cash balances increases aggregate demand, thereby boosting nominal GDP. Because velocity in our example is 4—that is, the dollar is spent, on average, four times per year—nominal GDP rises from \$400 billion to \$440 billion. At that higher nominal GDP, the money supply of \$110 billion equals the amount of money desired (\$440 billion/4 = \$110 billion), and equilibrium is reestablished.

The \$10 billion increase in the money supply thus eventually increases nominal GDP by \$40 billion. Spending on goods, services, and assets expands until nominal GDP has gone up enough to restore the original 4-to-1 equilibrium relationship between nominal GDP and the money supply.

Note that the relationship GDP/M defines V. A stable relationship between nominal GDP and M means a stable V. And a change in M causes a proportionate change in nominal GDP. Thus, monetarists say that changes in the money supply have a predictable effect on nominal GDP ($= P \times Q$). An increase in M increases P or Q, or some combination of both; a decrease in M reduces P or Q, or some combination of both.

WORKED PROBLEMS

W37.1
Equation of exchange

Monetary Causes of Instability Monetarists say that inappropriate monetary policy is the single most important cause of macroeconomic instability. An increase in the money supply directly increases aggregate demand. Under conditions of full employment, that rise in aggregate demand raises the price level. For a time, higher prices cause firms to increase their real output, and the rate of unemployment falls below its natural rate. But once nominal wages rise to reflect the higher prices and thus to restore real wages, real output moves back to its full-employment level and the unemployment rate returns to its natural rate. The inappropriate increase in the money supply leads to inflation, together with instability of real output and employment.

Conversely, a decrease in the money supply reduces aggregate demand. Real output temporarily falls, and the unemployment rate rises above its natural rate. Eventually, nominal wages fall and real output returns to its full-employment level. The inappropriate decline in the money supply leads to deflation, together with instability of real GDP and employment.

The contrast between mainstream macroeconomics and monetarism on the causes of instability thus comes into sharp focus. Mainstream economists view the instability of investment as the main cause of the economy's instability. They see monetary policy as a stabilizing factor. Changes in the money supply raise or lower interest rates as needed, smooth out swings in investment, and thus reduce macroeconomic instability. In contrast, monetarists view changes in the money supply as the main cause of instability in the economy. For example, they say that the Great Depression occurred largely because the Fed allowed the money supply to fall by roughly one-third during that period. According to Milton Friedman, a prominent monetarist,

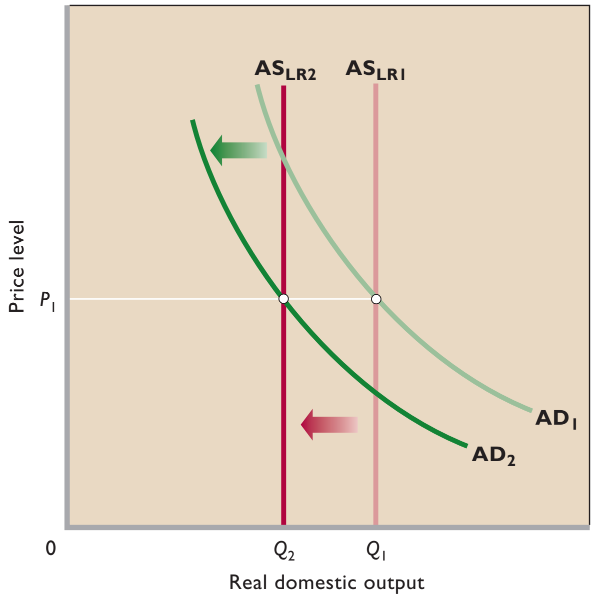
> And [the money supply] fell not because there were no willing borrowers—not because the horse would not drink. It fell because the Federal Reserve System forced or permitted a sharp reduction in the [money supply], because it failed to exercise the responsibilities assigned to it in the Federal Reserve Act to provide liquidity to the banking system. The Great Contraction is tragic testimony to the power of monetary policy—not, as Keynes and so many of his contemporaries believed, evidence of its impotence.[1]

Real-Business-Cycle View

A third modern view of the cause of macroeconomic instability is that business cycles are caused by real factors that affect aggregate supply rather than by monetary, or spending, factors that cause fluctuations in aggregate demand. In the **real-business-cycle theory,** business fluctuations result from significant changes in technology and resource availability. Those changes affect productivity and thus the long-run growth trend of aggregate supply.

An example focusing on recession will clarify this thinking. Suppose productivity (output per worker) declines sharply because of a large increase in oil prices, which makes it prohibitively expensive to operate certain types of machinery. That decline in productivity implies a reduction in the economy's ability to produce real output. The result would be a decrease in the economy's long-run aggregate supply curve, as represented by the leftward shift from AS_{LR1} to AS_{LR2} in Figure 37.1.

FIGURE 37.1 The real-business-cycle theory. In the real-business-cycle theory, a decline in resource availability shifts the nation's long-run aggregate supply curve to the left from AS_{LR1} to AS_{LR2}. The decline in real output from Q_1 to Q_2, in turn, reduces money demand (less is needed) and money supply (fewer loans are taken out) such that aggregate demand shifts leftward from AD_1 to AD_2. The result is a recession in which the price level remains constant.



As real output falls from Q_1 to Q_2, the public needs less money to buy the reduced volume of goods and services. So the demand for money falls. Moreover, the slowdown in business activity means that businesses need to borrow less from banks, reducing the part of the money supply created by banks through their lending. Thus, the supply of money also falls. In this controversial scenario, changes in the supply of money respond to changes in the demand for money. The decline in the money supply then reduces aggregate demand, as from AD_1 to AD_2 in Figure 37.1. The outcome is a decline in real output from Q_1 to Q_2, with no change in the price level.

Conversely, a large increase in aggregate supply (not shown) caused by, say, major innovations in the production process would shift the long-run aggregate supply curve rightward. Real output would increase, and money demand and money supply would both increase. Aggregate demand would shift rightward by an amount equal to the rightward shift of long-run aggregate supply. Real output would increase, without driving up the price level.

[1]Milton Friedman, *The Optimum Quantity of Money and Other Essays* (Chicago: Aldine, 1969), p. 97.

Conclusion: In the real-business-cycle theory, macro instability arises on the aggregate supply side of the economy, not on the aggregate demand side as mainstream economists and monetarists usually claim.

Coordination Failures

A fourth and final modern view of macroeconomic instability relates to so-called **coordination failures.** Such failures occur when people fail to reach a mutually beneficial equilibrium because they lack a way to coordinate their actions.

CONSIDER THIS . . .

Too Much Money?

The severe recession of 2007–2009 fits the mainstream view that recessions are caused by AD shocks rather than the real-business-cycle view that they are caused by AS shocks. Recall that the recession began with a financial crisis that caused significant reductions in investment spending and consumption spending and therefore reduced aggregate demand. As inventories increased, businesses further curtailed their investment spending, and aggregate demand, output, and income dropped even more.

Economists with monetarist leanings, however, cite monetary factors as the main cause of the financial crisis that led to the initial declines in aggregate demand. They argue that the Federal Reserve flooded the economy with too much money and held interest rates too low for too long in promoting recovery from the 2001 recession. In this line of reasoning, the excess money and low interest rates contributed to the bubble in the housing market. When that bubble burst, the resulting loan defaults set in motion the forces that produced the declines in AD and therefore the recession.

All economists agree that the bursting of the housing bubble led to the recession. And too-loose monetary policy may have contributed to the bubble. But economists also cite the role of the very large international flows of foreign savings into the United States during this period. These inflows drove down interest rates and helped to fuel the housing bubble. Other factors such as "pass-the-risk" lending practices, government policies promoting home ownership, and poorly designed and enforced financial regulations certainly came into play. The mainstream view is that causes of the financial crisis and recession are numerous and complex.

Noneconomic Example Consider first a noneconomic example. Suppose you hear that some people might be getting together at the last minute for an informal party at a nearby beach. But because of a chance of rain, there is some doubt about whether people will actually come out. You make a cell phone call or two to try to get a read on what others are thinking, and then base your decision on that limited information. If you expect others to be there, you will decide to go. If you expect that no one will go, you will decide to stay home. There are several possible equilibrium outcomes, depending on the mix of people's expectations. Let's consider just two. If each person assumes that all the others will go to the party, all will go. The party will occur and presumably everyone will have a good time. But if each person assumes that everyone else will stay home, all will stay home and there will be no party. When the party does not take place, even though all would be better off if it did take place, a coordination failure has occurred.

Macroeconomic Example Now let's apply this example to macroeconomic instability, specifically recession. Suppose that individual firms and households expect other firms and consumers to cut back their investment and consumption spending. As a result, each firm and household will anticipate a reduction of aggregate demand. Firms therefore will reduce their own investment spending since they will anticipate that their future production capacity will be excessive. Households will also reduce their own spending (increase their saving) because they anticipate that they will experience reduced work hours, possible layoffs, and falling incomes in the future.

Aggregate demand will indeed decline and the economy will indeed experience a recession in response to what amounts to a self-fulfilling prophecy. Moreover, the economy will stay at a below-full-employment level of output because, once there, producers and households have no individual incentive to increase spending. If all producers and households would agree to increase their investment and consumption spending simultaneously, then aggregate demand would rise, and real output and real income would increase. Each producer and each consumer would be better off. However, this outcome does not occur because there is no mechanism for firms and households to agree on such a joint spending increase.

In this case, the economy is stuck in an *unemployment equilibrium* because of a coordination failure. With a different set of expectations, a coordination failure might leave the economy in an *inflation equilibrium*. In this view, the economy has a number of such potential equilibrium positions, some good and some bad, depending on people's mix of expectations. Macroeconomic instability, then, reflects the movement of the economy from one such equilibrium position to another as expectations change.

QUICK REVIEW 37.1

- Mainstream economists say that macroeconomic instability usually stems from swings in investment spending and, occasionally, from adverse aggregate supply shocks.
- Monetarists view the economy through the equation of exchange ($MV = PQ$). If velocity V is stable, changes in the money supply M lead directly to changes in nominal GDP ($P \times Q$). For monetarists, changes in M caused by inappropriate monetary policy are the single most important cause of macroeconomic instability.
- In the real-business-cycle theory, significant changes in "real" factors such as technology, resource availability, and productivity change the economy's long-run aggregate supply, causing macroeconomic instability.
- Macroeconomic instability can result from coordination failures—less-than-optimal equilibrium positions that occur because businesses and households lack a way to coordinate their actions.

Does the Economy "Self-Correct"?

LO37.2 Discuss why new classical economists believe the economy will "self-correct" from aggregate demand and aggregate supply shocks.

Just as there are disputes over the causes of macroeconomic instability, there are disputes over whether or not the economy will correct itself when instability does occur. And economists also disagree on how long it will take for any such self-correction to take place.

New Classical View of Self-Correction

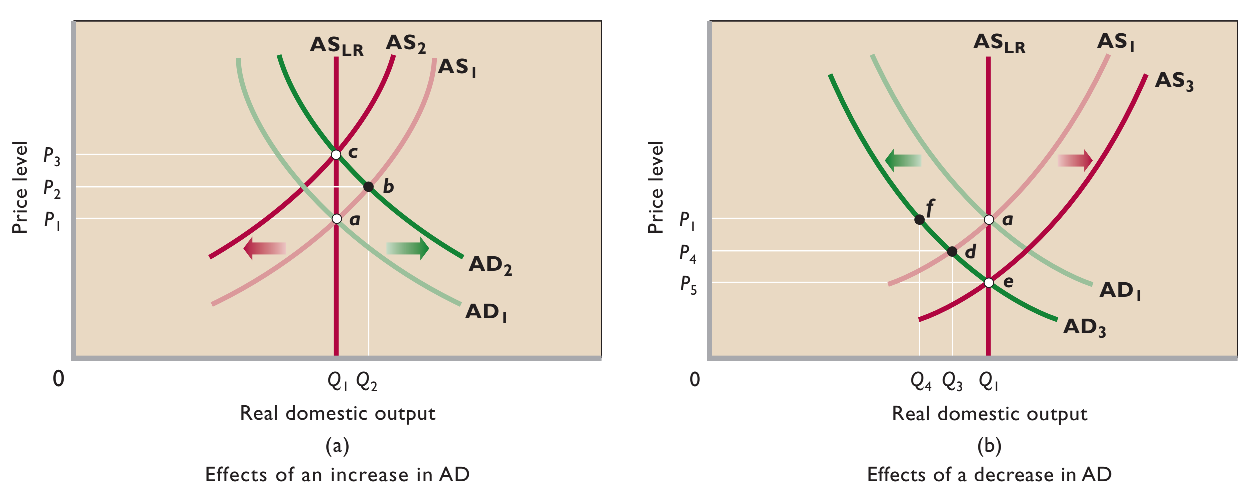
New classical economists tend to be either monetarists or adherents of **rational expectations theory:** the idea that businesses, consumers, and workers expect changes in policies or circumstances to have certain effects on the economy and, in pursuing their own self-interest, take actions to make sure those changes affect them as little as possible. The **new classical economics** holds that when the economy occasionally diverges from its full-employment output, internal mechanisms within the economy will automatically move it back to that output. Policymakers should stand back and let the automatic correction occur, rather than engage in active fiscal and monetary policy. This perspective is often associated with the vertical long-run Phillips Curve, which we discussed in Chapter 36. But we will analyze it here using the extended AD-AS model that was also developed in Chapter 36.

ORIGIN OF THE IDEA

O37.3
Rational expectations theory

Graphical Analysis Figure 37.2a relates the new classical analysis to the question of self-correction. Specifically, an increase in aggregate demand, say, from AD_1 to AD_2, moves

FIGURE 37.2 **New classical view of self-correction.** (a) An unanticipated increase in aggregate demand from AD_1 to AD_2 first moves the economy from *a* to *b*. The economy then self-corrects to *c*. An anticipated increase in aggregate demand moves the economy directly from *a* to *c*. (b) An unanticipated decrease in aggregate demand from AD_1 to AD_3 moves the economy from *a* to *d*. The economy then self-corrects to *e*. An anticipated decrease in aggregate demand moves the economy directly from *a* to *e*. (Mainstream economists, however, say that if the price level remains at P_1, the economy will move from *a* to *f*, and even if the price level falls to P_4, the economy may remain at *d* because of downward wage inflexibility.)



the economy upward along its short-run aggregate supply curve AS_1 from *a* to *b*. The price level rises and real output increases. With the economy producing beyond potential output, high resource demand drives up the prices of labor and other productive inputs. Per-unit production costs increase and the short-run aggregate supply curve shifts leftward, eventually from AS_1 to AS_2. The economy moves from *b* to *c*, and real output returns to its full-employment level, Q_1. This level of output is dictated by the economy's vertical long-run aggregate supply curve, AS_{LR}.

Conversely, a decrease in aggregate demand from AD_1 to AD_3 in Figure 37.2b first moves the economy downward along its short-run aggregate supply curve AS_1 from point *a* to *d*. The price level declines, as does the level of real output. With the economy producing below potential output, low resource demand drives down the prices of labor and other productive inputs. When that happens, per-unit production costs decline and the short-run aggregate supply curve shifts to the right, eventually from AS_1 to AS_3. The economy moves to *e*, where it again achieves its full-employment level, Q_1. As in Figure 37.2a, the economy in Figure 37.2b has automatically self-corrected to its full-employment output and its natural rate of unemployment.

Speed of Adjustment There is some disagreement among new classical economists on how long it will take for self-correction to occur. Monetarists usually hold the *adaptive expectations* view that people form their expectations on the basis of present realities and only gradually change their expectations as experience unfolds. This means that the shifts in the short-run aggregate supply curves shown in Figure 37.2 may not occur for 2 or 3 years or even longer. Other new classical economists, however, accept the rational expectations assumption that workers anticipate some future outcomes before they occur. When price-level changes are fully anticipated, adjustments of nominal wages are very quick or even instantaneous. Let's see why.

Although several new theories incorporate rational expectations, our interest here is the new classical version of the rational expectations theory (hereafter, RET). RET is based on two assumptions:

- People behave rationally, gathering and intelligently processing information to form expectations about things that are economically important to them. They adjust those expectations quickly as new developments affecting future economic outcomes occur. Where there is adequate information, people's beliefs about future economic outcomes accurately reflect the likelihood that those outcomes will occur. For example, if it is clear that a certain policy will cause inflation, people will recognize that fact and adjust their economic behavior in anticipation of inflation.
- RET economists assume that all product and resource markets are highly competitive and that prices and wages are flexible both upward and downward. But the RET economists go further, assuming that new information is quickly (in some cases, instantaneously) taken into account in the demand and supply curves of such markets. The upshot is that equilibrium prices and quantities adjust rapidly to unforeseen events—say, technological change or aggregate supply shocks. They adjust instantaneously to events that have known outcomes—for example, changes in fiscal or monetary policy.

Unanticipated Price-Level Changes The implication of RET is not only that the economy is self-correcting but that self-correction occurs quickly. In this thinking, unanticipated changes in the price level—so-called **price-level surprises**—*do* cause temporary changes in real output. Suppose, for example, that an unanticipated increase in foreign demand for U.S. goods increases U.S. aggregate demand from AD_1 to AD_2 in Figure 37.2a. The immediate result is an unexpected increase in the price level from P_1 to P_2.

But now an interesting question arises. If wages and prices are flexible, as assumed in RET, why doesn't the higher price level immediately cause wages and other input prices to rise, such that there is no increase in real output at all? Why does the economy temporarily move from point *a* to *b* along AS_1? In RET, firms increase output from Q_1 to Q_2 because of misperceptions about rising prices of their own products relative to the prices of other products (and to the prices of labor). They mistakenly think the higher prices of their own products have resulted from increased demand for those products relative to the demands for other products. Expecting higher profits, they increase their own production. But in fact *all* prices, including the price of labor (nominal wages), are rising because of the general increase in aggregate demand. Once firms see that *all* prices and wages are rising, they decrease their production to previous levels.

In terms of Figure 37.2a, the increase in nominal wages shifts the short-run aggregate supply curve leftward, ultimately from AS_1 to AS_2, and the economy moves from *b* to *c*. Thus, the increase in real output caused by the price-level surprise corrects itself.

The same analysis in reverse applies to an unanticipated price-level decrease. In the economy represented by Figure 37.2b, firms misperceive that the prices of their own products are falling due to decreases in the demand for those products relative to other products. They incorrectly anticipate declines in profit and cut production. As a result of their collective actions, real output in the economy falls. Once firms see what is really happening—that *all* prices and wages are dropping—they increase their output to prior levels. The short-run aggregate supply curve in Figure 37.2b shifts rightward from AS_1 to AS_3, and the economy "self-corrects" by moving from *d* to *e*.

Fully Anticipated Price-Level Changes In RET, fully *anticipated* price-level changes do not change real output, even for short periods. In Figure 37.2a, again consider the increase in aggregate demand from AD_1 to AD_2. Businesses immediately recognize that the higher prices being paid for their products are part of the inflation they had anticipated. They understand that the same forces that are causing the inflation result in higher nominal wages, leaving their profits unchanged. The economy therefore moves directly from *a* to *c*. The price level rises as expected, and output remains at its full-employment level Q_1.

Similarly, a fully *anticipated* price-level decrease will leave real output unchanged. Firms conclude that nominal wages are declining by the same percentage amount as the declining price level, leaving profits unchanged. The economy represented by Figure 37.2b therefore moves directly from *a* to *e*. Deflation occurs, but the economy continues to produce its full-employment output Q_1. The anticipated decline in aggregate demand causes no change in real output.

Mainstream View of Self-Correction

Almost all economists acknowledge that the new classical economists have made significant contributions to the theory of aggregate supply. In fact, mainstream economists have incorporated some aspects of RET into their own models. However, most economists strongly disagree with RET on the question of downward price and wage flexibility. While the stock market, foreign exchange market, and certain commodity markets experience day-to-day or minute-to-minute price changes, including price declines, that is not true of many product markets and most labor markets. There is ample evidence, say mainstream economists, that many prices and wages are inflexible downward for long periods. As a result, it may take years for the economy to move from recession back to full-employment output, unless it gets help from fiscal and monetary policy.

Graphical Analysis To understand this mainstream view, again examine Figure 37.2b. Suppose aggregate demand declines from AD_1 to AD_3 because of a significant decline in investment spending. If prices are sticky downward for a while and therefore the price level is temporarily stuck at P_1, the economy will not move from *a* to *d* to *e*, as suggested by RET. Instead, the economy will move from *a* to *f*, as if it were moving along the white horizontal P_1 price line between those two points. Real output will decline from its full-employment level, Q_1, to the recessionary level, Q_4.

But let's assume that large amounts of unsold inventories eventually cause the price level to fall to P_4. Will this lead to the decline in nominal wages needed to shift aggregate supply from AS_1 to AS_3, as suggested by the new classical economists? "Highly unlikely" say mainstream economists. Even more so than prices, nominal wages tend to be inflexible downward. If nominal wages do not decline in response to the decline in the price level, then the short-run aggregate supply curve will not shift rightward. The self-correction mechanism assumed by RET and new classical economists will break down. Instead, the economy will remain at *d*, experiencing less-than-full-employment output and a high rate of unemployment.

Downward Wage Inflexibility In Chapter 30 we discussed several reasons why firms may not be able to, or may not want to, lower nominal wages. Firms may not be able to cut wages because of wage contracts and the legal minimum wage. And firms may not want to lower wages if they fear potential problems with morale, effort, and efficiency.

While contracts are thought to be the main cause of wage rigidity, so-called efficiency wages and insider-outsider relationships also may play a role. Let's explore both.

Efficiency Wage Theory Recall from Chapter 30 that an **efficiency wage** is a wage that minimizes the firm's labor cost per unit of output. Normally, we would think that the market wage is the efficiency wage since it is the lowest wage at which a firm can obtain a particular type of labor. But where the cost of supervising workers is high or where worker turnover is great, firms may discover that paying a wage that is higher than the market wage will lower their wage cost per unit of output.

Example: Suppose a firm's workers, on average, produce 8 units of output at a \$9 market wage but 10 units of output at a \$10 above-market wage. The efficiency wage is \$10, not the \$9 market wage. At the \$10 wage, the labor cost per unit of output is only \$1 (= \$10 wage/10 units of output), compared with \$1.12 (= \$9 wage/8 units of output) at the \$9 wage.

How can a higher wage result in greater efficiency?

- ***Greater work effort*** The above-market wage, in effect, raises the cost to workers of losing their jobs as a result of poor performance. Because workers have a strong incentive to retain their relatively high-paying jobs, they are more likely to provide greater work effort. Looked at differently, workers are more reluctant to shirk (neglect or avoid work) because the higher wage makes job loss more costly to them. Consequently, the above-market wage can be the efficient wage; it can enhance worker productivity so much that the higher wage more than pays for itself.

- ***Lower supervision costs*** With less incentive among workers to shirk, the firm needs fewer supervisory personnel to monitor work performance. This, too, can lower the firm's overall wage cost per unit of output.
- ***Reduced job turnover*** The above-market pay discourages workers from voluntarily leaving their jobs. The lower turnover rate reduces the firm's cost of hiring and training workers. It also gives the firm a more experienced, more productive workforce.

The key implication for macroeconomic instability is that efficiency wages add to the downward inflexibility of wages. Firms that pay efficiency wages will be reluctant to cut wages when aggregate demand declines, since such cuts may encourage shirking, require more supervisory personnel, and increase turnover. In other words, wage cuts that reduce productivity and raise per-unit labor costs are self-defeating.

ORIGIN OF THE IDEA

O37.4
Efficiency wages

Insider-Outsider Relationships Other economists theorize that downward wage inflexibility may relate to relationships between "insiders" and "outsiders." Insiders are workers who retain employment even during recession. Outsiders are workers who have been laid off from a firm and unemployed workers who would like to work at that firm.

When recession produces layoffs and widespread unemployment, we might expect outsiders to offer to work for less than the current wage rate, in effect, bidding down wage rates. We also might expect firms to hire such workers to reduce their costs. But, according to the **insider-outsider theory,** outsiders may not be able to underbid existing wages because employers may view the nonwage cost of hiring them to be prohibitive. Employers might fear that insiders would view acceptance of such underbidding as undermining years of effort to increase wages or, worse, as "stealing" jobs. So insiders may refuse to cooperate with new workers who have undercut their pay. Where teamwork is critical for production, such lack of cooperation will reduce overall productivity and thereby lower the firms' profits.

Even if firms are willing to employ outsiders at less than the current wage, those workers might refuse to work for less than the existing wage. To do so might invite harassment from the insiders whose pay they have undercut. Thus, outsiders may remain unemployed, relying on past saving, unemployment compensation, and other social programs to make ends meet.

As in the efficiency wage theory, the insider-outsider theory implies that wages will be inflexible downward when aggregate demand declines. Self-correction may eventually occur but not nearly as rapidly as the new classical economists contend.

QUICK REVIEW 37.2

- New classical economists believe that the economy "self-corrects" when unanticipated events divert it from its full-employment level of real output.
- In RET, unanticipated price-level changes cause changes in real output in the short run but not in the long run.
- According to RET, market participants immediately change their actions in response to anticipated price-level changes such that no change in real output occurs.
- Mainstream economists say that the economy can get mired in recession for several months or more because of downward price and wage inflexibility.
- Sources of downward wage inflexibility include contracts, efficiency wages, and insider-outsider relationships.

Rules or Discretion?

LO37.3 Identify and describe the variations of the debate over "rules" versus "discretion" in conducting stabilization policy.

These different views on the causes of instability and on the speed of self-correction have led to vigorous debate on macro policy. Should the government adhere to policy rules that prohibit it from causing instability in an economy that is otherwise stable? Or should it use discretionary fiscal and monetary policy, when needed, to stabilize a sometimes-unstable economy?

In Support of Policy Rules

Monetarists and other new classical economists believe policy rules would reduce instability in the economy. They believe that such rules would prevent government from trying

CONSIDER THIS . . .

On the Road Again

Economist Abba Lerner (1903–1982) likened the economy to an automobile traveling down a road that had traffic barriers on each side. The problem was that the car had no steering wheel. It would hit one barrier, causing the car to veer to the opposite side of the road. There it would hit the other barrier, which in turn would send it careening to the opposite side. To avoid such careening in the form of business cycles, said Lerner, society must equip the economy with a steering wheel. Discretionary fiscal and monetary policy would enable government to steer the economy safely between the problems of recession and demand-pull inflation.

Economist Milton Friedman (1912–2006) modified Lerner's analogy, giving it a different meaning. He said that the economic vehicle does not need a skillful driver who is continuously turning the wheel to adjust to the unexpected irregularities of the route. Instead, the economy needs a way to prohibit the monetary passenger in the back seat from occasionally leaning over and giving the steering wheel a jerk that sends the car off the road. According to Friedman, the car will travel down the road just fine unless the Federal Reserve destabilizes it.

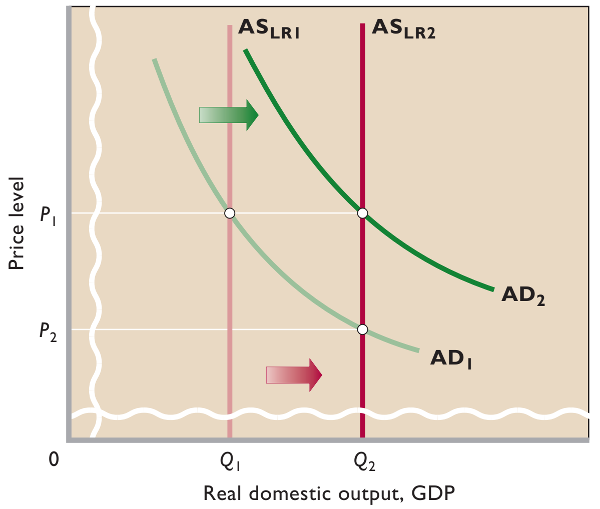
Lerner's analogy implied an internally unstable economy that needs steering through discretionary government stabilization policy. Friedman's modification of the analogy implied a generally stable economy that is destabilized by inappropriate monetary policy by the Federal Reserve. For Lerner, stability required active use of fiscal and monetary policy. For Friedman, macroeconomic stability required a monetary rule forcing the Federal Reserve to increase the money supply at a set, steady annual rate.*

*In his later years, Friedman softened his call for a monetary rule, acknowledging that the Fed had become much more skillful at keeping the rate of inflation in check through prudent monetary policy.

to "manage" aggregate demand. That would be a desirable trend because, in their view, such management is misguided and thus is likely to *cause* more instability than it cures.

Monetary Rule Since inappropriate monetary policy is the major source of macroeconomic instability, say monetarists, the enactment of a **monetary rule** would make sense.

FIGURE 37.3 Rationale for a monetary rule. A monetary rule that required the Fed to increase the money supply at an annual rate linked to the long-run increase in potential GDP would shift aggregate demand rightward, as from AD_1 to AD_2, at the same pace as the shift in long-run aggregate supply, here, AS_{LR1} to AS_{LR2}. Thus the economy would experience growth without inflation or deflation.



One such rule would be a requirement that the Fed expand the money supply each year at the same annual rate as the typical growth of the economy's production capacity. That fixed-rate expansion of the money supply would occur whatever the state of the economy. The Fed's sole monetary role would be to use its tools (open-market operations, the discount rate, interest on reserves, and reserve requirements) to ensure that the nation's money supply grew steadily by, say, 3 to 5 percent a year. According to Milton Friedman,

> Such a rule . . . would eliminate . . . the major cause of instability in the economy—the capricious and unpredictable impact of countercyclical monetary policy. As long as the money supply grows at a constant rate each year, be it 3, 4, or 5 percent, any decline into recession will be temporary. The liquidity provided by a constantly growing money supply will cause aggregate demand to expand. Similarly, if the supply of money does not rise at a more than average rate, any inflationary increase in spending will burn itself out for lack of fuel.[2]

Figure 37.3 illustrates the rationale for a monetary rule. Suppose the economy represented there is operating at its full-employment real output, Q_1. Also suppose the nation's long-run aggregate supply curve shifts rightward, as from AS_{LR1} to AS_{LR2}, each year, signifying the average annual increase in potential real output. As you saw in earlier chapters, such annual increases in "potential GDP"

[2]As quoted in Lawrence S. Ritter and William L. Silber, *Money*, 5th ed. (New York: Basic Books, 1984), pp. 141–142.

result from added resources, improved resources, and improved technology.

Monetarists argue that a monetary rule would tie increases in the money supply to the typical rightward shift of long-run aggregate supply. In view of the direct link between changes in the money supply and aggregate demand, this would ensure that the AD curve would shift rightward, as from AD_1 to AD_2, each year. As a result, while GDP increases from Q_1 to Q_2, the price level would remain constant at P_1. A monetary rule, then, would promote steady growth of real output along with price stability.

Generally, rational expectations economists also support a monetary rule. They conclude that an expansionary or restrictive monetary policy would alter the rate of inflation but not real output. Suppose, for example, the Fed implements an easy money policy to reduce interest rates, expand investment spending, and boost real GDP. On the basis of past experience and economic knowledge, the public would anticipate that this policy is inflationary and would take protective actions. Workers would press for higher nominal wages; firms would raise their product prices; and lenders would lift their nominal interest rates on loans.

All these responses are designed to prevent inflation from having adverse effects on the real income of workers, businesses, and lenders. But collectively they would immediately raise wage and price levels. So the increase in aggregate demand brought about by the expansionary monetary policy would be completely dissipated in higher prices and wages. Real output and employment would not be increased by the easy money policy.

In this view, the combination of rational expectations and instantaneous market adjustments dooms discretionary monetary policy to ineffectiveness. If discretionary monetary policy produces only inflation (or deflation), say the RET economists, then it makes sense to limit the Fed's discretion and to require that Congress enact a monetary rule consistent with price stability at all times.

In recent decades, the call for a Friedman-type monetary rule has faded. Some economists who tend to favor monetary rules have advocated **inflation targeting,** under which the Fed would be required to announce a targeted band of inflation rates, say, 1 to 2 percent, for some future period such as the following 2 years. It would then be expected to use its monetary policy tools to keep inflation rates within that range. If it did not hit the inflation target, it would have to explain why it failed.

Strictly interpreted, inflation targeting would focus the Fed's attention nearly exclusively on controlling inflation and deflation, rather than on counteracting business fluctuations. Proponents of inflation targeting generally believe the economy will have fewer, shorter, and less severe business cycles if the Fed adheres to the rule "Set a known inflation goal and achieve it."

We discussed another modern monetary rule—the Taylor rule—in Chapter 34 on monetary policy. This rule specifies how the Fed should alter the federal funds rate under differing economic circumstances. We discuss this rule in more depth in this chapter's Last Word.

Balanced Budget Monetarists and new classical economists question the effectiveness of fiscal policy. At the extreme, a few of them favor a constitutional amendment requiring that the federal government balance its budget annually. Others simply suggest that government be "passive" in its fiscal policy, not intentionally creating budget deficits or surpluses. They believe that deficits and surpluses caused by recession or inflationary expansion will eventually correct themselves as the economy self-corrects to its full-employment output.

Monetarists are particularly strong in their opposition to expansionary fiscal policy. They believe that the deficit spending accompanying such a policy has a strong tendency to "crowd out" private investment. Suppose government runs a budget deficit by printing and selling U.S. securities—that is, by borrowing from the public. By engaging in such borrowing, the government is competing with private businesses for funds. The borrowing increases the demand for money, which then raises the interest rate and crowds out a substantial amount of private investment that would otherwise have been profitable. The net effect of a budget deficit on aggregate demand therefore is unpredictable and, at best, modest.

RET economists reject discretionary fiscal policy for the same reason they reject active monetary policy: They don't think it works. Business and labor will immediately adjust their behavior in anticipation of the price-level effects of a change in fiscal policy. The economy will move directly to the anticipated new price level. Like monetary policy, say the RET theorists, fiscal policy can move the economy along its vertical long-run aggregate supply curve. But because its effects on inflation are fully anticipated, fiscal policy cannot alter real GDP even in the short run. The best course of action for government is to balance its budget.

In Defense of Discretionary Stabilization Policy

Mainstream economists oppose both a strict monetary rule and a balanced-budget requirement. They believe that monetary policy and fiscal policy are important tools for achieving and maintaining full employment, price stability, and economic growth.

Discretionary Monetary Policy In supporting discretionary monetary policy, mainstream economists argue that the rationale for the Friedman monetary rule is flawed. While there is indeed a close relationship between the money supply and nominal GDP over long periods, in shorter periods this relationship breaks down. The reason is that the velocity of money has proved to be more variable and unpredictable than monetarists contend. Arguing that velocity is variable both cyclically and over time, mainstream economists contend that a constant annual rate of increase in the money supply might not eliminate fluctuations in aggregate demand. In terms of the equation of exchange, a steady rise of M does not guarantee a steady expansion of aggregate demand because V—the rate at which money is spent—can change.

Look again at Figure 37.3, in which we demonstrated the monetary rule: Expand the money supply annually by a fixed percentage, regardless of the state of the economy. During the period in question, optimistic business expectations might create a boom in investment spending and thus shift the aggregate demand curve to some location to the right of AD_2. (You may want to pencil in a new AD curve, labeling it AD_3.) The price level would then rise above P_1; that is, demand-pull inflation would occur. In this case, the monetary rule will not accomplish its goal of maintaining price stability. Mainstream economists say that the Fed can use a restrictive monetary policy to reduce the excessive investment spending and thereby hold the rightward shift of aggregate demand to AD_2, thus avoiding inflation.

Similarly, suppose instead that investment declines because of pessimistic business expectations. Aggregate demand will then increase by some amount less than the increase from AD_1 to AD_2 in Figure 37.3. Again, the monetary rule fails the stability test: The price level sinks below P_1 (deflation occurs). Or if the price level is inflexible downward at P_1, the economy will not achieve its full-employment output (unemployment rises). An expansionary monetary policy can help avoid each outcome.

Mainstream economists quip that the trouble with the monetary rule is that it tells the policymaker, "Don't do something; just stand there."

Discretionary Fiscal Policy Mainstream economists support the use of fiscal policy to keep recessions from deepening or to keep mild inflation from becoming severe inflation. They recognize the possibility of crowding out but do not think it is a serious problem when business borrowing is depressed, as is usually the case in recession. Because politicians can abuse fiscal policy, most economists feel that it should be held in reserve for situations where monetary policy appears to be ineffective or working too slowly.

As indicated earlier, mainstream economists oppose requirements to balance the budget annually. Tax revenues fall sharply during recessions and rise briskly during periods of demand-pull inflation. Therefore, a law or a constitutional amendment mandating an annually balanced budget would require that the government increase tax rates and reduce government spending during recession and reduce tax rates and increase government spending during economic booms. The first set of actions would worsen recession, and the second set would fuel inflation.

Policy Successes

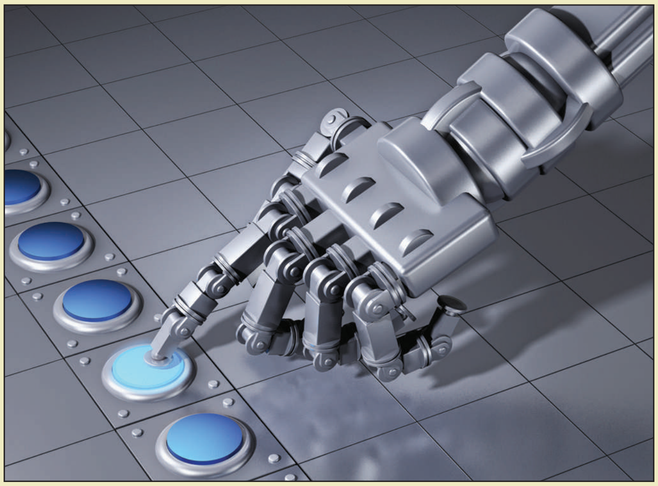
Finally, mainstream economists point out several specific policy successes in the past four decades:

- A tight money policy dropped inflation from 13.5 percent in 1980 to 3.2 percent in 1983.
- An expansionary fiscal policy reduced the unemployment rate from 9.7 percent in 1982 to 5.5 percent in 1988.
- An easy money policy helped the economy recover from the 1990–1991 recession.
- Judicious tightening of monetary policy in the mid-1990s, and then again in the late 1990s, helped the economy remain on a noninflationary, full-employment growth path.
- In late 2001 and 2002, expansionary fiscal and monetary policy helped the economy recover from a series of economic blows, including the collapse of numerous Internet start-up firms; a severe decline in investment spending; the impacts of the terrorist attacks of September 11, 2001; and a precipitous decline in stock values.
- In 2004 and 2005 the Fed tempered continued expansionary fiscal policy by increasing the federal funds rate in $\frac{1}{4}$ percentage-point increments from 1 percent to 4.25 percent. The economy expanded briskly in those years, while inflation stayed in check. The mild inflation was particularly impressive because the average price of a barrel of crude oil rose from \$24 in 2002 to \$55 in 2005. The Fed's further increases in interest rates to 5.25 percent in 2006 also kept inflation mild that year and the next despite continued strong growth and oil reaching \$99 per barrel in late 2007.
- In 2007 the Fed vigorously responded to a crisis in the mortgage market by ensuring monetary liquidity

LAST WORD

The Taylor Rule: Could a Robot Replace Ben Bernanke?

Macroeconomist John Taylor of Stanford University Calls for a New Monetary Rule That Would Institutionalize Appropriate Fed Policy Responses to Changes in Real Output and Inflation.



In our discussion of rules versus discretion, "rules" were associated with a *passive* monetary policy—one in which the monetary rule required that the Fed expand the money supply at a fixed annual rate regardless of the state of the economy. "Discretion," on the other hand, was associated with an *active* monetary policy in which the Fed changed interest rates in response to actual or anticipated changes in the economy.

Economist John Taylor has put a new twist on the rules-versus-discretion debate by suggesting a hybrid policy rule that dictates the precise active monetary actions the Fed should take when changes in the economy occur. You first encountered this Taylor rule in our discussion of monetary policy in Chapter 34. The **Taylor rule** combines traditional monetarism, with its emphasis on a monetary rule, and the more mainstream view that active monetary policy is a useful tool for taming inflation and limiting recession. Unlike the Friedman monetary rule, the Taylor rule holds, for example, that monetary policy should respond to changes in both real GDP and inflation, not simply inflation. The key adjustment instrument is the interest rate, not the money supply.

The Taylor rule builds on the belief held by many economists that central banks are willing to tolerate a small positive rate of inflation if doing so will help the economy to produce at potential output. The Taylor rule assumes that the Fed has a 2 percent "target rate of inflation" that it is willing to tolerate and that the Fed follows three rules when setting its target for the federal funds rate (the rate of interest that commercial banks with excess reserves charge on overnight loans to banks that wish to borrow reserves in order to meet their reserve requirements). The three rules are:

- When real GDP equals potential GDP and inflation is at its target rate of 2 percent, the federal funds target rate should be 4 percent, implying a *real* federal funds rate of 2 percent (= 4 percent nominal federal funds rate *minus* 2 percent inflation rate).
- For each 1 percent increase of real GDP above potential GDP, the Fed should raise the *real* federal funds rate by $\frac{1}{2}$ percentage point.
- For each 1 percent increase in the inflation rate above its 2 percent target rate, the Fed should raise the *real* federal funds rate by $\frac{1}{2}$ percentage point.*

Taylor has neither suggested nor implied that a robot, programmed with the Taylor rule, should replace Ben Bernanke, chairman of the Federal Reserve System. The Fed's discretion to override the rule (or "contingency plan for policy") would be retained, but the Fed would have to explain why its policies diverged from the rule. So the rule would remove the "mystery" associated with monetary policy and increase the Fed's accountability. Also, says Taylor, if used consistently, the rule would enable market participants to predict Fed behavior, and this would increase Fed credibility and reduce uncertainty.

Critics of the Taylor rule acknowledge that it is closer in tune with actual Fed countercyclical policy than is the Friedman monetary rule. They also acknowledge that adherence to the Taylor rule would have raised interest rates more rapidly during the two years prior to the mortgage debt crisis of 2007. Those higher interest rates might have reduced the excessive mortgage borrowing that eventually led to the crisis. But critics of the Taylor rule say that it is unwise to limit the Fed's future discretion in adjusting interest rates as it deems necessary. They note that the Fed has successfully kept inflation in check over the past two decades and has used monetary policy creatively and decisively when necessary during financial and economic emergencies. In view of the overall success, critics conclude that a mechanical monetary rule is certainly unnecessary and might even be detrimental.

*John Taylor, *Inflation, Unemployment, and Monetary Policy* (Cambridge, Mass.: MIT Press, 1998), pp. 44–47.

in the banking system. Besides aggressively lowering the federal funds rate from 5.25 percent in the summer of 2007 to just 2 percent in April 2008, the Fed greatly increased the reserves of the banking system by creating and using its term auction facility. It also undertook other lender-of-last-resort actions to unfreeze credit and prevent a possible collapse of the entire financial system. During the severe recession of 2007–2009, the Fed lowered the federal funds rate further, holding it in the 0 to 0.25 percent range to break the downward slide of the economy and help promote economic recovery.

QUICK REVIEW 37.3

- Mainstream economists disagree with monetarist and rational expectations economists as to whether the Fed should use rules or utilize discretion.
- Monetarist and rational expectations economists oppose discretion because they believe that discretion adds more instability to the business cycle than it cures.
- Mainstream economists oppose strict monetary rules and defend both monetary and fiscal policy discretion because they believe that both theory and evidence suggest that discretionary policies are helpful in achieving full employment, price stability, and economic growth.

Summary of Alternative Views

LO37.4 Summarize the fundamental ideas and policy implications of mainstream macroeconomics, monetarism, and rational expectations theory.

In Table 37.1 we summarize the fundamental ideas and policy implications of three macroeconomic theories: mainstream macroeconomics, monetarism, and rational expectations theory. Note that we have broadly defined new classical economics to include both monetarism and the rational expectations theory since both adhere to the view that the economy tends automatically to achieve equilibrium at its full-employment output.

These different perspectives have obliged mainstream economists to rethink some of their fundamental principles and to revise many of their positions. Although considerable disagreement remains, mainstream macroeconomists agree with monetarists that "money matters" and that excessive growth of the money supply is the major cause of long-lasting, rapid inflation. They also agree with RET proponents and theorists of coordination failures that expectations matter. If government can create expectations of price stability, full employment, and economic growth, households and firms will tend to act in ways to make them happen. In short, thanks to ongoing challenges to conventional wisdom, macroeconomics continues to evolve.

TABLE 37.1 Summary of Alternative Macroeconomic Views

		New Classical Economics	
Issue	Mainstream Macroeconomics	Monetarism	Rational Expectations
View of the private economy	Potentially unstable	Stable in long run at natural rate of unemployment	Stable in long run at natural rate of unemployment
Cause of the observed instability of the private economy	Investment plans unequal to saving plans (changes in AD); AS shocks	Inappropriate monetary policy	Unanticipated AD and AS shocks in the short run
Assumptions about short-run price and wage stickiness	Both prices and wages stuck in the immediate short run; in the short run, wages sticky while prices inflexible downward but flexible upward	Prices flexible upward and downward in the short run; wages sticky in the short run	Prices and wages flexible both upward and downward in the short run
Appropriate macro policies	Active fiscal and monetary policy	Monetary rule	Monetary rule
How changes in the money supply affect the economy	By changing the interest rate, which changes investment and real GDP	By directly changing AD, which changes GDP	No effect on output because price-level changes are anticipated
View of the velocity of money	Unstable	Stable	No consensus
How fiscal policy affects the economy	Changes AD and GDP via the multiplier process	No effect unless money supply changes	No effect because price-level changes are anticipated
View of cost-push inflation	Possible (AS shock)	Impossible in the long run in the absence of excessive money supply growth	Impossible in the long run in the absence of excessive money supply growth

SUMMARY

LO37.1 Describe alternative perspectives on the causes of macroeconomic instability, including the views of mainstream economists, monetarists, real-business-cycle advocates, and proponents of coordination failures.

The mainstream view is that macro instability is caused by a combination of price stickiness and shocks to aggregate demand or aggregate supply. With prices inflexible in the shorter run, changes in aggregate demand or short-run aggregate supply result in changes in output and employment. In the long run when both input and output prices are fully flexible, the economy will produce at potential output.

Monetarism focuses on the equation of exchange: $MV = PQ$. Because velocity is thought to be stable, changes in M create changes in nominal GDP ($= PQ$). Monetarists believe that the most significant cause of macroeconomic instability has been inappropriate monetary policy. Rapid increases in M cause inflation; insufficient growth of M causes recession. In this view, a major cause of the Great Depression was inappropriate monetary policy, which allowed the money supply to decline by roughly one-third.

Real-business-cycle theory views changes in resource availability and technology (real factors), which alter productivity, as the main causes of macroeconomic instability. In this theory, shifts of the economy's long-run aggregate supply curve change real output. In turn, money demand and money supply change, shifting the aggregate demand curve in the same direction as the initial change in long-run aggregate supply. Real output thus can change without a change in the price level.

A coordination failure is said to occur when people lack a way to coordinate their actions in order to achieve a mutually beneficial equilibrium. Depending on people's expectations, the economy can come to rest at either a good equilibrium (noninflationary full-employment output) or a bad equilibrium (less-than-full-employment output or demand-pull inflation). A bad equilibrium is a result of a coordination failure.

The rational expectations theory rests on two assumptions: (1) With sufficient information, people's beliefs about future economic outcomes accurately reflect the likelihood that those outcomes will occur; and (2) markets are highly competitive, and prices and wages are flexible both upward and downward.

LO37.2 Discuss why new classical economists believe the economy will "self-correct" from aggregate demand and aggregate supply shocks.

New classical economists (monetarists and rational expectations theorists) see the economy as automatically correcting itself when disturbed from its full-employment level of real output. In RET, unanticipated changes in aggregate demand change the price level, and in the short run this leads firms to change output. But once the firms realize that all prices are changing (including nominal wages) as part of general inflation or deflation, they restore their output to the previous level. Anticipated changes in aggregate demand produce only changes in the price level, not changes in real output.

Mainstream economists reject the new classical view that all prices and wages are flexible downward. They contend that nominal wages, in particular, are inflexible downward because of several factors, including labor contracts, efficiency wages, and insider-outsider relationships. This means that declines in aggregate demand lower real output, not only wages and prices.

LO37.3 Identify and describe the variations of the debate over "rules" versus "discretion" in conducting stabilization policy.

Monetarist and rational expectations economists say the Fed should adhere to some form of policy rule, rather than rely exclusively on discretion. The Friedman rule would direct the Fed to increase the money supply at a fixed annual rate equal to the long-run growth of potential GDP. An alternative approach—inflation targeting—would direct the Fed to establish a targeted range of inflation rates, say, 1 to 2 percent, and focus monetary policy on meeting that goal. They also support maintaining a "neutral" fiscal policy, as opposed to using discretionary fiscal policy to create budget deficits or budget surpluses. A few monetarists and rational expectations economists favor a constitutional amendment requiring that the federal government balance its budget annually.

Mainstream economists oppose strict monetary rules and a balanced-budget requirement, and defend discretionary monetary and fiscal policies. They say that both theory and evidence suggest that such policies are helpful in achieving full employment, price stability, and economic growth.

LO37.4 Summarize the fundamental ideas and policy implications of mainstream macroeconomics, monetarism, and rational expectations theory.

Macroeconomics continues to evolve because of the debate among the three main schools of economic thought: the mainstream, monetarism, and rational expectations.

Mainstream economics believes the economy to be potentially unstable and prone to business cycles due to sticky prices and wages interacting with economic shocks. The mainstream believes in active monetary and fiscal policy that can change AD through the multiplier process and that changes in the money supply affect the economy by changing the interest rate and investment. The mainstream believes the velocity of money to be unstable and holds that cost-push inflation exists and is caused by AS shocks.

Monetarism believes the economy to be fundamentally stable unless inappropriate monetary policies are applied. It believes wages and prices are fully flexible at all time horizons, allowing the economy to adjust on its own back to equilibrium. It argues that business cycles will only occur if monetary policy is either overly tight or overly loose. To prevent either, monetarists support the application of a *monetary rule* to guide monetary policy. Monetarism views the velocity of money as stable and believes cost-push inflation to be impossible in the long run in the absence of excessive money supply growth. Monetarists believe that fiscal policy cannot affect AD or GDP unless there are accompanying changes in monetary policy. On the other hand, monetary policy can shift AD and thereby affect GDP.

The rational expectations camp believes the economy to be stable in the long run at the natural rate of unemployment due to its assumption that prices are flexible at all time horizons. But over shorter time horizons, recessions and booms can happen due to unexpected AD or AS shocks. Rational expectations proponents support a monetary rule in order to guide expectations, so that monetary policy changes will never be a surprise that could shift the economy away from its equilibrium. They also agree with monetarists in the belief that cost-push inflation should be impossible without excessive money supply growth. Finally, rational expectations economists believe that anticipated fiscal and monetary policy changes will have no effect on GDP because they will lead only to price-level changes. For either to be effective in shifting GDP in the short run, they must come as a surprise.

TERMS AND CONCEPTS

monetarism
equation of exchange
velocity
real-business-cycle theory
coordination failures
rational expectations theory
new classical economics
price-level surprises
efficiency wage
insider-outsider theory
monetary rule
inflation targeting
Taylor rule

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. According to mainstream economists, what is the usual cause of macroeconomic instability? What role does the spending-income multiplier play in creating instability? How might adverse aggregate supply factors cause instability, according to mainstream economists? **LO37.1**
2. What is an efficiency wage? How might payment of an above-market wage reduce shirking by employees and reduce worker turnover? How might efficiency wages contribute to downward wage inflexibility, at least for a time, when aggregate demand declines? **LO37.1**
3. How might relationships between so-called insiders and outsiders contribute to downward wage inflexibility? **LO37.1**
4. Briefly describe the difference between a so-called real business cycle and a more traditional "spending" business cycle. **LO37.1**
5. Craig and Kris were walking directly toward each other in a congested store aisle. Craig moved to his left to avoid Kris, and at the same time Kris moved to his right to avoid Craig. They bumped into each other. What concept does this example illustrate? How does this idea relate to macroeconomic instability? **LO37.1**
6. State and explain the basic equation of monetarism. What is the major cause of macroeconomic instability, as viewed by monetarists? **LO37.1**
7. Use the equation of exchange to explain the rationale for a monetary rule. Why will such a rule run into trouble if V unexpectedly falls because of, say, a drop in investment spending by businesses? **LO37.1**
8. Explain the difference between "active" discretionary fiscal policy advocated by mainstream economists and "passive" fiscal policy advocated by new classical economists. Explain: "The problem with a balanced-budget amendment is that it would, in a sense, require active fiscal policy—but in the wrong direction—as the economy slides into recession." **LO37.3**
9. You have just been elected president of the United States, and the present chairperson of the Federal Reserve Board has resigned. You need to appoint a new person to this position, as well as a person to chair your Council of Economic Advisers. Using Table 37.1 and your knowledge of macroeconomics, identify the views on macro theory and policy you would want your appointees to hold. Remember, the economic health of the entire nation—and your chances for reelection—may depend on your selection. **LO37.4**
10. **LAST WORD** Compare and contrast the Taylor rule for monetary policy with the older, simpler monetary rule advocated by Milton Friedman.

REVIEW QUESTIONS

1. If prices are sticky and the number of dollars of gross investment unexpectedly increases, the __________ curve will shift __________. **LO37.1**
   a. AD; right.
   b. AD; left.
   c. AS; right.
   d. AS; left.
2. First, imagine that both input and output prices are fixed in the economy. What does the aggregate supply curve look like? If AD decreases in this situation, what will happen to equilibrium output and the price level? Next, imagine that input prices are fixed, but output prices are flexible. What does the aggregate supply curve look like? In this case, if AD decreases, what will happen to equilibrium output and the price level? Finally, if both input and output prices are fully flexible, what does the aggregate supply curve look like? In this case, if AD decreases, what will happen to equilibrium output and the price level? (To check your answers, review Figures 30.3, 30.4, and 30.5 in Chapter 30). **LO37.1**
3. Suppose that the money supply is $1 trillion and money velocity is 4. Then the equation of exchange would predict nominal GDP to be: **LO37.1**
   a. $1 trillion.
   b. $4 trillion.
   c. $5 trillion.
   d. $8 trillion.
4. If the money supply fell by 10 percent, a monetarist would expect nominal GDP to __________. **LO37.1**
   a. Rise.
   b. Fall.
   c. Stay the same.
5. An economy is producing at full employment when AD unexpectedly shifts to the left. A new classical economist would assume that as the economy adjusted back to producing at full employment, the price level would __________. **LO37.2**
   a. Increase.
   b. Decrease.
   c. Stay the same.
6. Use an AD-AS graph to demonstrate and explain the price-level and real-output outcome of an anticipated decline in aggregate demand, as viewed by RET economists. (Assume that the economy initially is operating at its full-employment level of output.) Then demonstrate and explain on the same graph the outcome as viewed by mainstream economists. **LO37.2**
7. Place "MON," "RET," or "MAIN" beside the statements that most closely reflect monetarist, rational expectations, or mainstream views, respectively: **LO37.4**
   a. Anticipated changes in aggregate demand affect only the price level; they have no effect on real output.
   b. Downward wage inflexibility means that declines in aggregate demand can cause long-lasting recession.
   c. Changes in the money supply M increase PQ; at first only Q rises, because nominal wages are fixed, but once workers adapt their expectations to new realities, P rises and Q returns to its former level.
   d. Fiscal and monetary policies smooth out the business cycle.
   e. The Fed should increase the money supply at a fixed annual rate.

PROBLEMS

1. Suppose that the money supply and the nominal GDP for a hypothetical economy are $96 billion and $336 billion, respectively. What is the velocity of money? How will households and businesses react if the central bank reduces the money supply by $20 billion? By how much will nominal GDP have to fall to restore equilibrium, according to the monetarist perspective? **LO37.1**
2. Assume the following information for a hypothetical economy in year 1: money supply = $400 billion; long-term annual growth of potential GDP = 3 percent; velocity = 4. Assume that the banking system initially has no excess reserves and that the reserve requirement is 10 percent. Also suppose that velocity is constant and that the economy initially is operating at its full-employment real output. **LO37.1**
   a. What is the level of nominal GDP in year 1?
   b. Suppose the Fed adheres to a monetary rule through open-market operations. What amount of U.S. securities will it have to sell to, or buy from, banks or the public between years 1 and 2 to meet its monetary rule?

PART **ELEVEN**

INTERNATIONAL ECONOMICS

CHAPTER **38**

International Trade*

Learning Objective

LO38.1 **List and discuss several key facts about international trade.**

LO38.2 **Define comparative advantage, and demonstrate how specialization and trade add to a nation's output.**

LO38.3 **Describe how differences between world prices and domestic prices prompt exports and imports.**

LO38.4 **Analyze the economic effects of tariffs and quotas.**

LO38.5 **Analyze the validity of the most frequently presented arguments for protectionism.**

LO38.6 **Identify and explain the objectives of GATT, WTO, EU, eurozone, and NAFTA, and discuss offshoring and trade adjustment assistance.**

Backpackers in the wilderness like to think they are "leaving the world behind," but, like Atlas, they carry the world on their shoulders. Much of their equipment is imported—knives from Switzerland, rain gear from South Korea, cameras from Japan, aluminum pots from England, sleeping bags from China, and compasses from Finland. Moreover, they may have driven to the trailheads in Japanese-made Toyotas or German-made BMWs, sipping coffee from Brazil or snacking on bananas from Honduras.

International trade and the global economy affect all of us daily, whether we are hiking in the wilderness, driving our cars, buying groceries, or working at our jobs. We cannot "leave the world behind." We are enmeshed in a global web of economic relationships, such as trading goods and services, multinational corporations, cooperative

*Note to Instructors: If you prefer to cover international trade early in your course, you can assign this chapter at the end of either Part 1 or Part 2. This chapter builds on the introductory ideas of opportunity costs, supply and demand analysis, and economic efficiency but does not require an understanding of either market failures or government failures.

ventures among the world's firms, and ties among the world's financial markets.

The focus of this chapter is the trading of goods and services. Then in Chapter 39, we examine the U.S. balance of payments, exchange rates, and U.S. trade deficits. In Chapter 39W on our Web site, we look at the economics of developing nations.

Some Key Trade Facts

LO38.1 List and discuss several key facts about international trade.

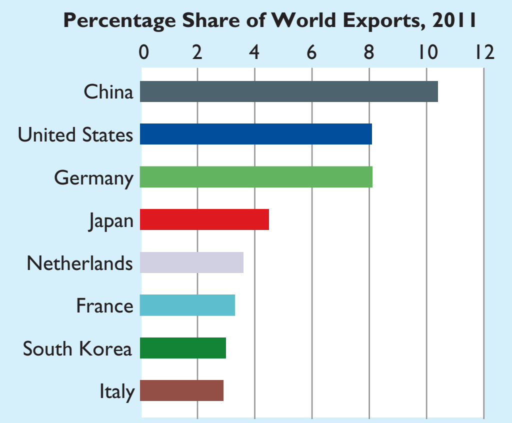
The following are several important facts relating to international trade.

- U.S. exports and imports have more than doubled as percentages of GDP since 1980.
- A *trade deficit* occurs when imports exceed exports. The United States has a trade deficit in goods. In 2012 U.S. imports of goods exceeded U.S. exports of goods by $735 billion.
- A *trade surplus* occurs when exports exceed imports. The United States has a trade surplus in services (such as air transportation services and financial services). In 2012 U.S. exports of services exceeded U.S. imports of services by $196 billion.
- Principal U.S. exports include chemicals, agricultural products, consumer durables, semiconductors, and aircraft; principal imports include petroleum, automobiles, metals, household appliances, and computers.
- As with other advanced industrial nations, the United States imports many goods that are in some of the same categories as the goods that it exports. Examples: automobiles, computers, chemicals, semiconductors, and telecommunications equipment.
- Canada is the United States' most important trading partner quantitatively. In 2012 about 20 percent of U.S. exported goods were sold to Canadians, who in turn provided 15 percent of imported U.S. goods.
- The United States has a sizable trade deficit with China. In 2012 it was $315 billion.
- The U.S. dependence on foreign oil is reflected in its trade with members of OPEC. In 2012 the United States imported $181 billion of goods (mainly oil) from OPEC members, while exporting $82 billion of goods to those countries.
- The United States leads the world in the combined volume of exports and imports, as measured in dollars. China, the United States, Germany, Japan, and the Netherlands were the top five exporters by dollar in 2012.
- Currently, the United States provides about 8.1 percent of the world's exports. (See Global Perspective 38.1.)
- Exports of goods and services (on a national income account basis) make up about 14 percent of total U.S. output. That percentage is much lower than the percentage in many other nations, including Canada, France, Germany, the Netherlands, and South Korea. (See Global Perspective 38.2.)
- China has become a major international trader, with an estimated $2.05 trillion of exports in 2012. Other Asian economies—including South Korea, Taiwan, and Singapore—are also active in international trade. Their combined exports exceed those of France, Britain, or Italy.

GLOBAL PERSPECTIVE 38.1

Shares of World Exports, Selected Nations

China has the largest share of world exports, followed by Germany and the United States. The eight largest export nations account for about 43.9 percent of world exports.

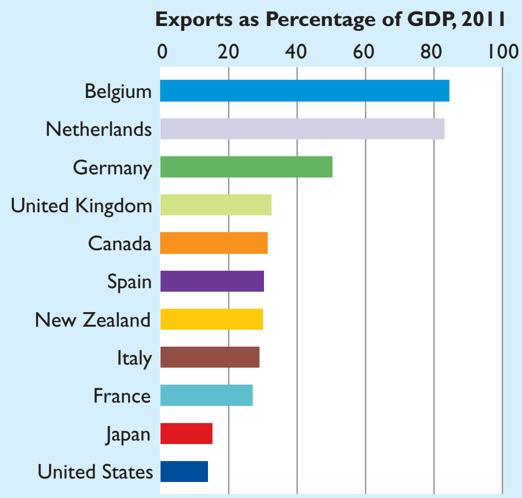


Source: International Trade Statistics, 2012, WTO Publications.

GLOBAL PERSPECTIVE 38.2

Exports of Goods and Services as a Percentage of GDP, Selected Countries

Although the United States is one of the world's largest exporters, as a percentage of GDP, its exports are quite low relative to many other countries.



Source: Derived by authors from *IMF International Financial Statistics*, 2012.

- International trade links world economies. Through trade, changes in economic conditions in one place on the globe can quickly affect other places.
- International trade is often at the center of debates over economic policy, both within the United States and internationally.

With this information in mind, let's turn to the economics of international trade.

The Economic Basis for Trade

LO38.2 Define comparative advantage, and demonstrate how specialization and trade add to a nation's output.

Sovereign nations, like individuals and the regions of a nation, can gain by specializing in the products they can produce with the greatest relative efficiency and by trading for the goods they cannot produce as efficiently. The simple answer to the question "Why do nations trade?" is "They trade because it is beneficial." The benefits that emerge relate to three underlying facts:

- The distribution of natural, human, and capital resources among nations is uneven; nations differ in their endowments of economic resources.
- Efficient production of various goods requires different technologies, and not all nations have the same level of technological expertise.
- Products are differentiated as to quality and other attributes, and some people may prefer certain goods imported from abroad rather than similar goods produced domestically.

To recognize the character and interaction of these three facts, think of China, which has abundant and inexpensive labor. As a result, China can produce efficiently (at low cost of other goods forgone) a variety of **labor-intensive goods,** such as textiles, electronics, apparel, toys, and sporting goods.

In contrast, Australia has vast amounts of land and can inexpensively produce such **land-intensive goods** as beef, wool, and meat. Mexico has the soil, tropical climate, rainfall, and ready supply of unskilled labor that allow for the efficient, low-cost production of vegetables. Industrially advanced economies such as the United States and Germany that have relatively large amounts of capital can inexpensively produce goods whose production requires much capital, including such **capital-intensive goods** as airplanes, automobiles, agricultural equipment, machinery, and chemicals.

Also, regardless of their resource intensities, nations can develop individual products that are in demand worldwide because of their special qualities. Examples: fashions from Italy, chocolates from Belgium, software from the United States, and watches from Switzerland.

The distribution of resources, technology, and product distinctiveness among nations is relatively stable in short time periods but certainly can change over time. When that distribution changes, the relative efficiency and success that nations have in producing and selling goods also change. For example, in the past several decades, South Korea has greatly expanded its stock of capital. Although South Korea was primarily an exporter of agricultural products and raw materials a half-century ago, it now exports large quantities of manufactured goods. Similarly, the new technologies that gave us synthetic fibers and synthetic rubber drastically altered the resource mix needed to produce fibers and rubber and changed the relative efficiency of nations in manufacturing them.

As national economies evolve, the size and quality of their labor forces may change, the volume and composition of their capital stocks may shift, new technologies may develop, and even the quality of land and the quantity of natural resources may be altered. As such changes take place, the relative efficiency with which a nation can produce specific goods will also change. As economists would say, comparative advantage can and does sometimes change.

Comparative Advantage

In an open economy (one with an international sector), a country produces more of certain goods (exports) and fewer of other goods (imports) than it would otherwise. Thus, the country shifts the use of labor and other productive resources toward export industries and away from import industries. For example, in the presence of international trade, the United States uses more resources to make commercial aircraft and to grow wheat and fewer resources to make television sets and sew clothes. So we ask: Do such shifts of resources make economic sense? Do they enhance U.S. total output and thus the U.S. standard of living?

The answers are affirmative. Specialization and international trade increase the productivity of U.S. resources and allow the United States to obtain greater total output than otherwise would be possible. These benefits are the result of exploiting both *absolute advantages* and *comparative advantages*. A country is said to have an *absolute advantage* over other producers of a product if it is the most efficient producer of that product (by which we mean that it can produce more output of that product from any given amount of resource inputs than can any other producer). A country is said to have a *comparative advantage* over other producers of a product if it can produce the product at a lower opportunity cost (by which we mean that it must forgo less output of alternative products when allocating productive resources to producing the product in question).

In 1776 Adam Smith used the concept of absolute advantage to argue for international specialization and trade. His point was that nations would be better off if each specialized in the production of those products in which it had an absolute advantage and was therefore the most efficient producer:

> It is the maxim of every prudent master of a family, never to attempt to make at home what it will cost him more to make than to buy. The taylor does not attempt to make his own shoes, but buys them of the shoemaker. The shoemaker does not attempt to make his own clothes, but employs a taylor. The farmer attempts to make neither the one nor the other, but employs those different artificers. . . .
>
> What is prudence in the conduct of every private family, can scarce be folly in that of a great kingdom. If a foreign country can supply us with a commodity cheaper than we can make it, better buy it of them with some part of the produce of our own industry, employed in a way in which we have some advantage.[1]

[1]Adam Smith, *The Wealth of Nations* (originally published, 1776; New York: Modern Library, 1937), p. 424.

CONSIDER THIS . . .

A CPA and a House Painter

Suppose that Madison, a certified public accountant (CPA), is a swifter painter than Mason, the professional painter she is thinking of hiring. Also assume that Madison can earn $50 per hour as an accountant but would have to pay Mason $15 per hour. And suppose that Madison would need 30 hours to paint her house but Mason would need 40 hours.

Should Madison take time from her accounting to paint her own house, or should she hire the painter? Madison's opportunity cost of painting her house is $1,500 (= 30 hours of sacrificed CPA time × $50 per CPA hour). The cost of hiring Mason is only $600 (= 40 hours of painting × $15 per hour of painting). Although Madison is better at both accounting and painting, she will get her house painted at lower cost by specializing in accounting and using some of her earnings from accounting to hire a house painter.

Similarly, Mason can reduce his cost of obtaining accounting services by specializing in painting and using some of his income to hire Madison to prepare his income tax forms. Suppose Mason would need 10 hours to prepare his tax return, while Madison could handle the task in 2 hours. Mason would sacrifice $150 of income (= 10 hours of painting time × $15 per hour) to do something he could hire Madison to do for $100 (= 2 hours of CPA time × $50 per CPA hour). By specializing in painting and hiring Madison to prepare his tax return, Mason lowers the cost of getting his tax return prepared.

We will see that what is true for our CPA and house painter is also true for nations. Specializing on the basis of comparative advantage enables nations to reduce the cost of obtaining the goods and services they desire.

In the early 1800s, David Ricardo extended Smith's idea by demonstrating that it is advantageous for a country to specialize and trade with another country even if it is more productive in all economic activities than that other country. Stated more formally, a nation does not need Smith's absolute advantage—total superiority in the efficiency with which it produces products—to benefit from specialization and trade. It needs only a comparative advantage.

The nearby Consider This box provides a simple, two-person illustration of Ricardo's principle of comparative advantage. Be sure to read it now because it will greatly help you understand the graphical analysis that follows.

QUICK REVIEW 38.1

- International trade enables nations to specialize, increase productivity, and increase the amount of output available for consumption.
- A country is said to have an *absolute advantage* over the other producers of a product if it can produce the product more efficiently, by which we mean that it can produce more of the product from any given amount of resource inputs than can any other producer.
- A country is said to have a *comparative advantage* over the other producers of a product if it can produce the product at a lower opportunity cost, by which we mean that it must forgo less of the output of alternative products when allocating resources to producing the product in question.

Two Isolated Nations

Our goal is to place the idea of comparative advantage into the context of trading nations. Our method is to build a simple model that relies on the familiar concepts of production possibilities curves. Suppose the world consists of just two nations, the United States and Mexico. Also for simplicity, suppose that the labor forces in the United States and Mexico are of equal size. Each nation can produce both beef and raw (unprocessed) vegetables but at different levels of economic efficiency. Suppose the U.S. and Mexican domestic production possibilities curves for beef and vegetables are those shown in Figure 38.1a and Figure 38.1b. Note three realities relating to the production possibilities curves in the two graphs:

- ***Constant costs*** The curves derive from the data in Table 38.1 and are drawn as straight lines, in contrast to the bowed-outward production possibilities frontiers we examined in Chapter 1. This means that we have replaced the law of increasing opportunity costs with the assumption of constant costs. This substitution simplifies our discussion but does not impair the validity of our analysis and conclusions. Later we will consider the effects of increasing opportunity costs.
- ***Different costs*** The production possibilities curves of the United States and Mexico reflect different resource mixes and differing levels of technology. Specifically, the differing slopes of the two curves reflect the numbers in the figures and reveal that the opportunity costs of producing beef and vegetables differ between the two nations.
- ***U.S. absolute advantage in both*** A producer (an individual, firm, or country) has an *absolute advantage* over another producer if it can produce more of a product than the other producer using the same amount of resources. Because of our convenient assumption that the U.S. and Mexican labor forces are the same size, the two production possibilities curves show that the United States has an absolute advantage in producing both products. If the United States and Mexico use their entire (equal-size) labor forces to produce either vegetables or beef, the United States can produce more of either than Mexico. The United States, using the same number of workers as Mexico, has greater production possibilities. So output per worker—labor productivity—in the United States exceeds that in Mexico in producing both products.

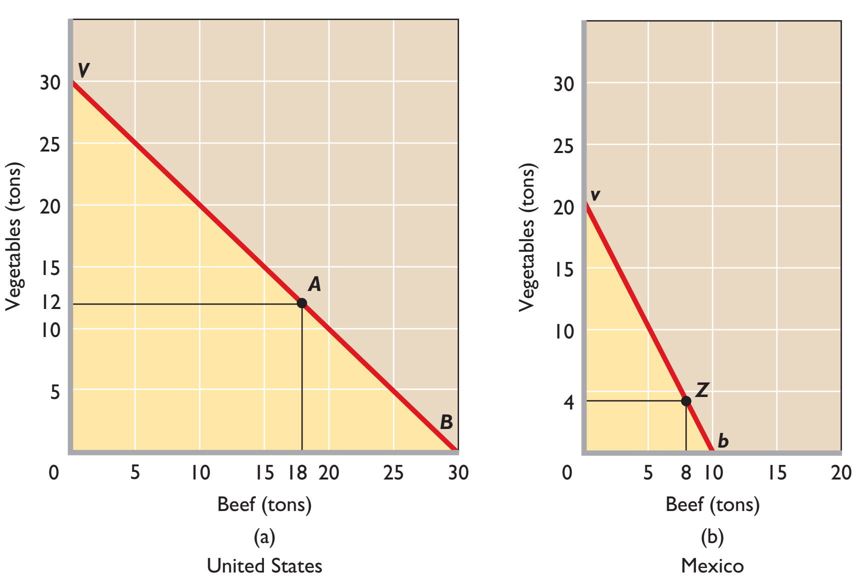


FIGURE 38.1 Production possibilities for the United States and Mexico. The two production possibilities curves show the combinations of vegetables and beef that the United States and Mexico can produce domestically. The curves for both countries are straight lines because we are assuming constant opportunity costs. (a) As reflected by the slope of *VB* in the left graph, the opportunity-cost ratio in the United States is 1 vegetables ≡ 1 beef. (b) The production possibilities curve *vb* in the right graph has a steeper slope, reflecting the different opportunity-cost ratio in Mexico of 2 vegetables ≡ 1 beef. The difference in the opportunity-cost ratios between the two countries defines their comparative advantages and is the basis for specialization and international trade.

TABLE 38.1 **International Specialization According to Comparative Advantage and the Gains from Trade**

Country	(1) Outputs before Specialization	(2) Outputs after Specialization	(3) Amounts Exported (−) and Imported (+)	(4) Outputs Available after Trade	(5) Gains from Specialization and Trade (4) − (1)
United States	18 beef	30 beef	−10 beef	20 beef	2 beef
	12 vegetables	0 vegetables	+15 vegetables	15 vegetables	3 vegetables
Mexico	8 beef	0 beef	+10 beef	10 beef	2 beef
	4 vegetables	20 vegetables	−15 vegetables	5 vegetables	1 vegetables

Opportunity-Cost Ratio in the United States In Figure 38.1a, with full employment, the United States will operate at some point on its production possibilities curve. On that curve, it can increase its output of beef from 0 tons to 30 tons by forgoing 30 tons of vegetables output. So the slope of the production possibilities curve is 1 (= 30 vegetables/30 beef), meaning that 1 ton of vegetables must be sacrificed for each extra ton of beef. In the United States the **opportunity-cost ratio** (domestic exchange ratio) for the two products is 1 ton of vegetables (V) for 1 ton of beef (B), or

United States: $1V \equiv 1B$ (The "≡" sign simply means "equivalent to.")

Within its borders, the United States can "exchange" a ton of vegetables from itself for a ton of beef from itself. Our constant-cost assumption means that this exchange or opportunity-cost relationship prevails for all possible moves from one point to another along the U.S. production possibilities curve.

Opportunity-Cost Ratio in Mexico Mexico's production possibilities curve in Figure 38.1b represents a different full-employment opportunity-cost ratio. In Mexico, 20 tons of vegetables must be given up to obtain 10 tons of beef. The slope of the production possibilities curve is 2 (= 20 vegetables/10 beef). This means that in Mexico the opportunity-cost ratio for the two goods is 2 tons of vegetables for 1 ton of beef, or

Mexico: $2V \equiv 1B$

Self-Sufficiency Output Mix If the United States and Mexico are isolated and self-sufficient, then each country must choose some output mix on its production possibilities curve. It will select the mix that provides the greatest total utility or satisfaction. Let's assume that combination point A in Figure 38.1a is the optimal mix in the United States. That is, society deems the combination of 18 tons of beef and 12 tons of vegetables preferable to any other combination of the goods available along the production possibilities curve. Suppose Mexico's optimal product mix is 8 tons of beef and 4 tons of vegetables, indicated by point Z in Figure 38.1b. These choices by the two countries are reflected in column 1 of Table 38.1.

Specializing Based on Comparative Advantage

A producer (an individual, firm, or nation) has a **comparative advantage** in producing a particular product if it can produce that product at a lower opportunity cost than other producers. Comparative advantage is the key determinant in whether or not nations can gain from specialization and trade. In fact, absolute advantage turns out to be irrelevant.

In our example, for instance, the United States has an absolute advantage over Mexico in producing both vegetables and beef. But it is still the case that the United States can gain from specialization and trade with Mexico. That is because what actually matters is whether the opportunity costs of producing the two products (beef and vegetables) differ in the two countries. If they do, then each nation will enjoy a comparative advantage in one of the products, meaning that it can produce that product at a lower opportunity cost than the other country. As a result, total output can increase if each country specializes in the production of the good in which it has the lower opportunity cost.

This idea is summarized in the **principle of comparative advantage,** which says that total output will be greatest when each good is produced by the nation that has the lowest domestic opportunity cost for producing that good. In our two-nation illustration, the United States has the lower domestic opportunity cost for beef; the United States must forgo only 1 ton of vegetables to produce 1 ton of beef, whereas Mexico must forgo 2 tons of vegetables for 1 ton of beef. The United States has a comparative (cost) advantage in beef and should specialize in beef production. The "world" (that is, the United States and Mexico) in our example would clearly not be economizing in the use of its resources if a high-cost producer (Mexico) produced a specific product (beef) when a low-cost producer (the United States) could have produced it. Having Mexico produce beef would mean that the world economy would have to give up more vegetables than is necessary to obtain a ton of beef.

Mexico has the lower domestic opportunity cost for vegetables. It must sacrifice only $\frac{1}{2}$ ton of beef to produce 1 ton of vegetables, while the United States must forgo 1 ton of beef to produce 1 ton of vegetables. Mexico has a comparative advantage in vegetables and should specialize in vegetable production. Again, the world would not be employing its resources economically if vegetables were produced by a high-cost producer (the United States) rather than by a low-cost producer (Mexico). If the United States produced vegetables, the world would be giving up more beef than necessary to obtain each ton of vegetables. Economizing requires that any particular good be produced by the nation having the lowest domestic opportunity cost or the nation having the comparative advantage for that good. The United States should produce beef, and Mexico should produce vegetables. The situation is summarized in Table 38.2.

A comparison of columns 1 and 2 in Table 38.1 verifies that specialized production enables the world to obtain more output from its fixed amount of resources. By specializing completely in beef, the United States can produce 30 tons of beef and no vegetables. Mexico, by specializing completely in vegetables, can produce 20 tons of vegetables and no beef. These figures exceed the yields generated without specialization: 26 tons of beef (= 18 in the United States + 8 in Mexico) and 16 tons of vegetables (= 12 in the United States + 4 in Mexico). As a result, the world ends up with 4 more tons of beef (= 30 tons − 26 tons) and 4 more tons of vegetables (= 20 tons − 16 tons) than it would if there were self-sufficiency and unspecialized production.

ORIGIN OF THE IDEA

O38.1
Absolute and comparative advantage

TABLE 38.2 **Comparative-Advantage Example: A Summary**

Beef	Vegetables
Mexico: Must give up 2 tons of vegetables to get 1 ton of beef.	***Mexico:*** Must give up $\frac{1}{2}$ ton of beef to get 1 ton of vegetables.
United States: Must give up 1 ton of vegetables to get 1 ton of beef.	***United States:*** Must give up 1 ton of beef to get 1 ton of vegetables.
Comparative advantage: United States	***Comparative advantage:*** Mexico

Terms of Trade

We have just seen that specialization in production will allow for the largest possible amounts of both beef and vegetables to be produced. But with each country specializing in the production of only one item, how will the vegetables that are all produced by Mexico and the beef that is all produced by the United States be divided between consumers in the two countries? The key turns out to be the **terms of trade,** the exchange ratio at which the United States and Mexico trade beef and vegetables.

Crucially, the terms of trade also establish whether each country will find it in its own better interest to bother specializing at all. This is because the terms of trade determine whether each country can "get a better deal" by specializing and trading than it could if it opted instead for self sufficiency. To see how this works, note that because $1B \equiv 1V$ (= $1V \equiv 1B$) in the United States, it must get more than 1 ton of vegetables for each 1 ton of beef exported; otherwise, it will not benefit from exporting beef in exchange for Mexican vegetables. The United States must get a better "price" (more vegetables) for its beef through international trade than it can get domestically; otherwise, no gain from trade exists and such trade will not occur.

Similarly, because $1B \equiv 2V$ (= $2V \equiv 1B$) in Mexico, Mexico must obtain 1 ton of beef by exporting less than 2 tons of vegetables to get it. Mexico must be able to pay a lower "price" for beef in the world market than it must pay domestically, or else it will not want to trade. The international exchange ratio or terms of trade must therefore lie somewhere between

$$1B \equiv 1V \text{ (United States' cost conditions)}$$

and

$$1B \equiv 2V \text{ (Mexico's cost conditions)}$$

Where between these limits will the world exchange ratio fall? The United States will prefer a rate close to $1B \equiv 2V$, say, $1B \equiv 1\frac{3}{4}V$. The United States wants to obtain as many vegetables as possible for each 1 ton of beef it exports. By contrast, Mexico wants a rate near $1B \equiv 1V$, say, $1B \equiv 1\frac{1}{4}V$. This is true because Mexico wants to export as few vegetables as possible for each 1 ton of beef it receives in exchange.

The actual exchange ratio depends on world supply and demand for the two products. If overall world demand for vegetables is weak relative to its supply and if the demand for beef is strong relative to its supply, the price of vegetables will be lower and the price of beef will be higher. The exchange ratio will settle nearer the $1B \equiv 2V$ figure the United States prefers. If overall world demand for vegetables is great relative to its supply and if the demand for beef is weak relative to its supply, the ratio will settle nearer the

KEY GRAPH

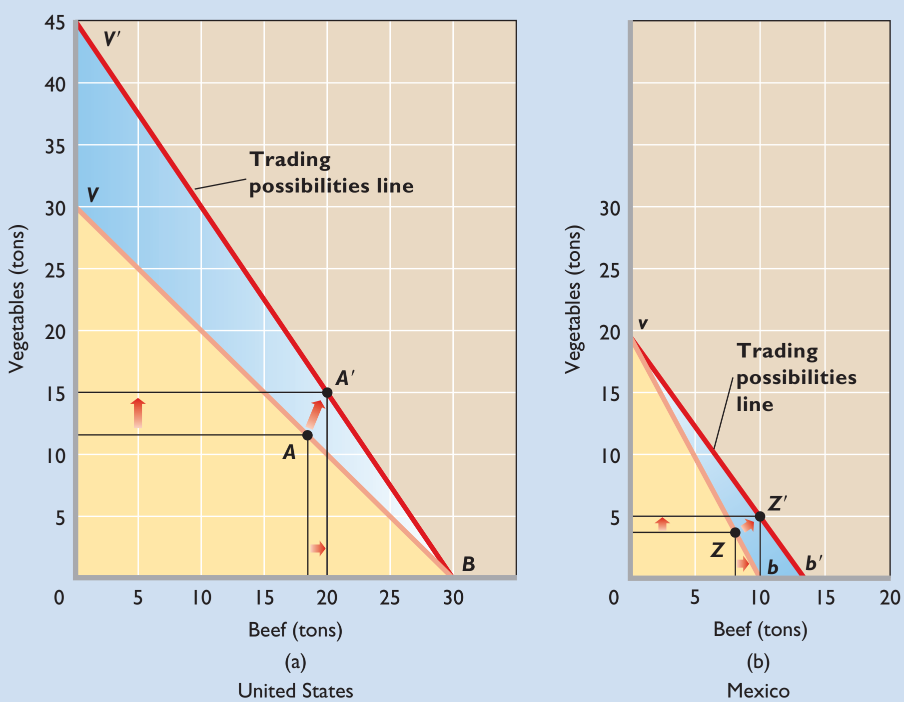


FIGURE 38.2 Trading possibilities lines and the gains from trade. As a result of specialization and trade, both the United States and Mexico can have higher levels of output than the levels attainable on their domestic production possibilities curves. (a) The United States can move from point *A* on its domestic production possibilities curve to, say, *A′* on its trading possibilities line. (b) Mexico can move from *Z* to *Z′*.

QUICK QUIZ FOR FIGURE 38.2

1. The production possibilities curves in graphs (a) and (b) imply:
- **a.** increasing domestic opportunity costs.
- **b.** decreasing domestic opportunity costs.
- **c.** constant domestic opportunity costs.
- **d.** first decreasing, then increasing, domestic opportunity costs.

2. Before specialization, the domestic opportunity cost of producing 1 unit of beef is:
- **a.** 1 unit of vegetables in both the United States and Mexico.
- **b.** 1 unit of vegetables in the United States and 2 units of vegetables in Mexico.
- **c.** 2 units of vegetables in the United States and 1 unit of vegetables in Mexico.
- **d.** 1 unit of vegetables in the United States and $\frac{1}{2}$ unit of vegetables in Mexico.

3. After specialization and international trade, the world output of beef and vegetables is:
- **a.** 20 tons of beef and 20 tons of vegetables.
- **b.** 45 tons of beef and 15 tons of vegetables.
- **c.** 30 tons of beef and 20 tons of vegetables.
- **d.** 10 tons of beef and 30 tons of vegetables.

4. After specialization and international trade:
- **a.** the United States can obtain units of vegetables at less cost than it could before trade.
- **b.** Mexico can obtain more than 20 tons of vegetables, if it so chooses.
- **c.** the United States no longer has a comparative advantage in producing beef.
- **d.** Mexico can benefit by prohibiting vegetables imports from the United States.

Answers: 1. c; 2. b; 3. c; 4. a

$1B \equiv 1V$ level favorable to Mexico. In this manner, the actual exchange ratio that is set by world supply and demand determines how the gains from international specialization and trade are divided between the two nations and, consequently, how the beef that is all produced in the United States and the vegetables that are all produced in Mexico get divided among consumers in the two countries. (We discuss equilibrium world prices later in this chapter.)

Gains from Trade

Suppose the international terms of trade are $1B \equiv 1\frac{1}{2}V$. The possibility of trading on these terms permits each nation to supplant its domestic production possibilities curve with a trading possibilities line (or curve), as shown in **Figure 38.2 (Key Graph).** Just as a production possibilities curve shows the amounts of these products that a full-employment economy can obtain by shifting resources from one to the other, a **trading possibilities line** shows the amounts of the two products that a nation can obtain by specializing in one product and trading for the other. The trading possibilities lines in Figure 38.2 reflect the assumption that both nations specialize on the basis of comparative advantage: The United States specializes completely in beef (at point *B* in Figure 38.2a), and Mexico specializes completely in vegetables (at point *v* in Figure 38.2b).

Improved Alternatives With specialization and trade, the United States is no longer constrained by its domestic production possibilities line, which requires it to give up 1 ton of beef for every 1 ton of vegetables it wants as it moves up its domestic production possibilities line from, say, point *B*. Instead, the United States, through trade with Mexico, can get $1\frac{1}{2}$ tons of vegetables for every ton of beef that it exports to Mexico, as long as Mexico has vegetables to export. Trading possibilities line BV' thus represents the $1B \equiv 1\frac{1}{2}V$ trading ratio.

Similarly, Mexico, starting at, say, point *v*, no longer has to move down its domestic production possibilities curve, giving up 2 tons of vegetables for each ton of beef it wants. It can now export just $1\frac{1}{2}$ tons of vegetables for each 1 ton of beef that it wants by moving down its trading possibilities line vb'.

Specialization and trade create a new exchange ratio between beef and vegetables, and that ratio is reflected in each nation's trading possibilities line. For both nations, this exchange ratio is superior to the unspecialized exchange ratio embodied in their respective production possibilities curves. By specializing in beef and trading for Mexico's vegetables, the United States can obtain more than 1 ton of vegetables for 1 ton of beef. By specializing in vegetables and trading for U.S. beef, Mexico can obtain 1 ton of beef for less than 2 tons of vegetables. In both cases, self-sufficiency is inefficient and therefore undesirable.

Greater Output By specializing on the basis of comparative advantage and by trading for goods that are produced in the nation with greater domestic efficiency, the United States and Mexico can achieve combinations of beef and vegetables beyond their own individual production possibilities curves. Specialization according to comparative advantage results in a more efficient allocation of world resources, and larger outputs of both products are therefore available to both nations.

Suppose that at the $1B \equiv 1\frac{1}{2}V$ terms of trade, the United States exports 10 tons of beef to Mexico and in return Mexico exports 15 tons of vegetables to the United States. How do the new quantities of beef and vegetables available to the two nations compare with the optimal product mixes that existed before specialization and trade? Point *A* in Figure 38.2a reminds us that the United States chose 18 tons of beef and 12 tons of vegetables originally. But by producing 30 tons of beef and no vegetables and by trading 10 tons of beef for 15 tons of vegetables, the United States can obtain 20 tons of beef and 15 tons of vegetables. This new, superior combination of beef and vegetables is indicated by point A' in Figure 38.2a. Compared with the no-trade amounts of 18 tons of beef and 12 tons of vegetables, the United States' **gains from trade** are 2 tons of beef and 3 tons of vegetables.

WORKED PROBLEMS

W38.1
Gains from trade

Similarly, recall that Mexico's optimal product mix was 4 tons of vegetables and 8 tons of beef (point *Z*) before specialization and trade. Now, after specializing in vegetables and trading for beef, Mexico can have 5 tons of vegetables and 10 tons of beef. It accomplishes that by producing 20 tons of vegetables and no beef and exporting 15 tons of its vegetables in exchange for 10 tons of American beef. This new position is indicated by point Z' in Figure 38.2b. Mexico's gains from trade are 1 ton of vegetables and 2 tons of beef.

Points A' and Z' in Figure 38.2 are superior economic positions to points *A* and *Z*. This fact is enormously important! We know that a nation can expand its production possibilities boundary by (1) expanding the quantity and improving the quality of its resources or (2) realizing technological progress. We have now established that international trade can enable a nation to circumvent the output constraint illustrated by its production possibilities curve. An economy can grow by expanding international trade. The outcome of international specialization and trade is equivalent to having more and better resources or discovering and implementing improved production techniques.

Table 38.1 summarizes the transactions and outcomes in our analysis. Please give it one final careful review.

QUICK REVIEW 38.2

- The principle of comparative advantage says that total world output will be greatest when each good is produced by the nation that has the lowest domestic opportunity cost.
- The rate at which countries can trade units of one product for units of another product is referred to as the terms of trade.
- A trading possibilities line shows the amounts of two products that a nation can obtain by specializing in the production of one product and then trading for the other.

CONSIDER THIS . . .

Misunderstanding the Gains from Trade

It is a common myth that the greatest benefit to be derived from international trade is greater domestic employment in the export sector. This suggests that exports are "good" because they increase domestic employment, whereas imports are "bad" because they deprive people of jobs at home. As we have demonstrated, the true benefit created by international trade is the overall increase in output available through specialization and exchange.

A nation does not need international trade to operate *on* its production possibilities curve. It can fully employ its resources, including labor, with or without international trade. International trade, however, enables a country to reach a point of consumption beyond its domestic production possibilities curve. The gain from trade to a nation is the extra output obtained from abroad—the imports obtained for less sacrifice of other goods than if they were produced at home.

Trade with Increasing Costs

To explain the basic principles underlying international trade, we simplified our analysis in several ways. For example, we limited discussion to two products and two nations. But multiproduct and multinational analysis yield the same conclusions. We also assumed constant opportunity costs (linear production possibilities curves), which is a more substantive simplification. Let's consider the effect of allowing increasing opportunity costs (concave-to-the-origin production possibilities curves) to enter the picture.

Suppose that the United States and Mexico initially are at positions on their concave production possibilities curves where their domestic cost ratios are $1B \equiv 1V$ and $1B \equiv 2V$, as they were in our constant-cost analysis. As before, comparative advantage indicates that the United States should specialize in beef and Mexico in vegetables. But now, as the United States begins to expand beef production, its cost of beef will rise; it will have to sacrifice more than 1 ton of vegetables to get 1 additional ton of beef. Resources are no longer perfectly substitutable between alternative uses, as the constant-cost assumption implied. Resources less and less suitable to beef production must be allocated to the U.S. beef industry in expanding beef output, and that means increasing costs—the sacrifice of larger and larger amounts of vegetables for each additional ton of beef.

Similarly, suppose that Mexico expands vegetable production starting from its $1B \equiv 2V$ cost ratio position. As production increases, it will find that its $1B \equiv 2V$ cost ratio begins to rise. Sacrificing 1 ton of beef will free resources that are capable of producing only something less than 2 tons of vegetables because those transferred resources are less suitable to vegetable production.

As the U.S. cost ratio falls from $1B \equiv 1V$ and the Mexican ratio rises from $1B \equiv 2V$, a point will be reached where the cost ratios are equal in the two nations, perhaps at $1B \equiv 1\frac{3}{4}V$. At this point the underlying basis for further specialization and trade—differing cost ratios—has disappeared, and further specialization is therefore uneconomical. And, most important, this point of equal cost ratios may be reached while the United States is still producing some vegetables along with its beef and Mexico is producing some beef along with its vegetables. The primary effect of increasing opportunity costs is less-than-complete specialization. For this reason, we often find domestically produced products competing directly against identical or similar imported products within a particular economy.

The Case for Free Trade

The case for free trade reduces to one compelling argument: Through free trade based on the principle of comparative advantage, the world economy can achieve a more efficient allocation of resources and a higher level of material well-being than it can without free trade.

Since the resource mixes and technological knowledge of the world's nations are all somewhat different, each nation can produce particular commodities at different real costs. Each nation should produce goods for which its domestic opportunity costs are lower than the domestic opportunity costs of other nations and exchange those goods for products for which its domestic opportunity costs are high relative to those of other nations. If each nation does this, the world will realize the advantages of geographic and human specialization. The world and each free-trading nation can obtain a larger real income from the fixed supplies of resources available to it.

Government trade barriers lessen or eliminate gains from specialization. If nations cannot trade freely, they must shift resources from efficient (low-cost) to inefficient (high-cost) uses to satisfy their diverse wants. A recent study suggests that the elimination of trade barriers since the Second World War has increased the income of the average U.S. household by at least $7,000 and perhaps by as much as $13,000. These income gains are recurring; they happen year after year.[2]

One side benefit of free trade is that it promotes competition and deters monopoly. The increased competition from foreign firms forces domestic firms to find and use the lowest-cost production techniques. It also compels them to be innovative with respect to both product quality and production methods, thereby contributing to economic growth. And free trade gives consumers a wider range of product choices. The reasons to favor free trade are the same as the reasons to endorse competition.

A second side benefit of free trade is that it links national interests and breaks down national animosities. Confronted with political disagreements, trading partners tend to negotiate rather than make war.

QUICK REVIEW 38.3

- International trade enables nations to specialize, increase productivity, and increase output available for consumption.
- Comparative advantage means total world output will be greatest when each good is produced by the nation that has the lowest domestic opportunity cost.
- Specialization is less than complete among nations because opportunity costs normally rise as any specific nation produces more of a particular good.

Supply and Demand Analysis of Exports and Imports

LO38.3 Describe how differences between world prices and domestic prices prompt exports and imports.

Supply and demand analysis reveals how equilibrium prices and quantities of exports and imports are determined. The amount of a good or a service a nation will export or import depends on differences between the equilibrium world price and the equilibrium domestic price. The interaction of *world* supply and demand determines the equilibrium **world price**—the price that equates the quantities supplied and demanded globally. *Domestic* supply and demand determine the equilibrium **domestic price**—the price that would prevail in a closed economy that does not engage in international trade. The domestic price equates quantity supplied and quantity demanded domestically.

[2]Scott C. Bradford, Paul L. E. Grieco, and Gary C. Hufbauer, "The Payoff to America from Globalization," *The World Economy*, July 2006, pp. 893–916.

In the absence of trade, the domestic prices in a closed economy may or may not equal the world equilibrium prices. When economies are opened for international trade, differences between world and domestic prices encourage exports or imports. To see how, consider the international effects of such price differences in a simple two-nation world, consisting of the United States and Canada, that are both producing aluminum. We assume there are no trade barriers, such as tariffs and quotas, and no international transportation costs.

Supply and Demand in the United States

Figure 38.3a shows the domestic supply curve S_d and the domestic demand curve D_d for aluminum in the United States, which for now is a closed economy. The intersection of S_d and D_d determines the equilibrium domestic price of $1 per pound and the equilibrium domestic quantity of 100 million pounds. Domestic suppliers produce 100 million pounds and sell them all at $1 a pound. So there are no domestic surpluses or shortages of aluminum.

But what if the U.S. economy were opened to trade and the world price of aluminum were above or below this $1 domestic price?

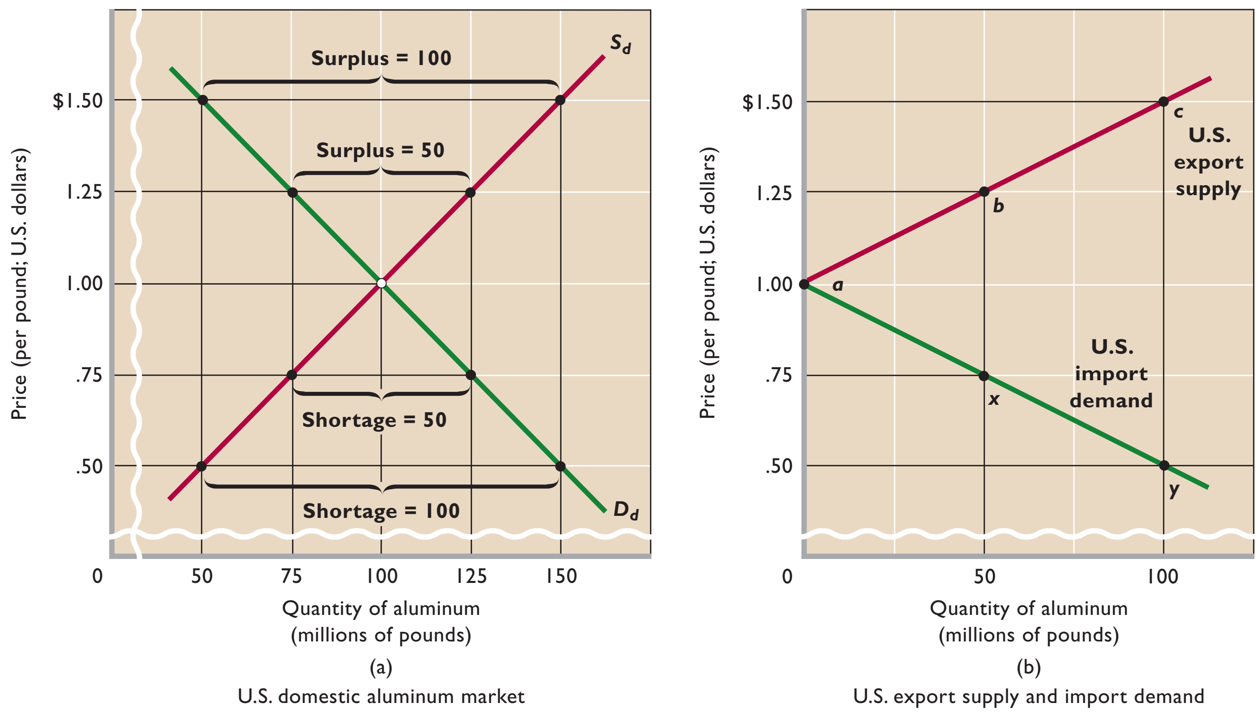
U.S. Export Supply If the aluminum price in the rest of the world (that is, Canada) exceeds $1, U.S. firms will produce more than 100 million pounds and will export the excess domestic output. First, consider a world price of $1.25. We see from the supply curve S_d that U.S. aluminum firms will produce 125 million pounds of aluminum at that price. The demand curve D_d tells us that the United States will purchase only 75 million pounds at $1.25. The outcome is a domestic surplus of 50 million pounds of aluminum. U.S. producers will export those 50 million pounds at the $1.25 world price.

What if the world price were $1.50? The supply curve shows that U.S. firms will produce 150 million pounds of aluminum, while the demand curve tells us that U.S. consumers will buy only 50 million pounds. So U.S. producers will export the domestic surplus of 100 million pounds.

Toward the top of Figure 38.3b we plot the domestic surpluses—the U.S. exports—that occur at world prices above the $1 domestic equilibrium price. When the world and domestic prices are equal (= $1), the quantity of exports supplied is zero (point *a*). There is no surplus of domestic output to export. But when the world price is $1.25, U.S. firms export 50 million pounds of surplus aluminum (point *b*). At a $1.50 world price, the domestic surplus of 100 million pounds is exported (point *c*).

The U.S. **export supply curve,** found by connecting points *a*, *b*, and *c*, shows the amount of aluminum U.S.

FIGURE 38.3 U.S. export supply and import demand. (a) Domestic supply S_d and demand D_d set the domestic equilibrium price of aluminum at \$1 per pound. At world prices above \$1 there are domestic surpluses of aluminum. At prices below \$1 there are domestic shortages. (b) Surpluses are exported (top curve), and shortages are met by importing aluminum (lower curve). The export supply curve shows the direct relationship between world prices and U.S. exports; the import demand curve portrays the inverse relationship between world prices and U.S. imports.



producers will export at each world price above \$1. This curve *slopes upward*, indicating a direct or positive relationship between the world price and the amount of U.S. exports. As world prices increase relative to domestic prices, U.S. exports rise.

U.S. Import Demand If the world price is below the domestic \$1 price, the United States will import aluminum. Consider a \$0.75 world price. The supply curve in Figure 38.3a reveals that at that price U.S. firms produce only 75 million pounds of aluminum. But the demand curve shows that the United States wants to buy 125 million pounds at that price. The result is a domestic shortage of 50 million pounds. To satisfy that shortage, the United States will import 50 million pounds of aluminum.

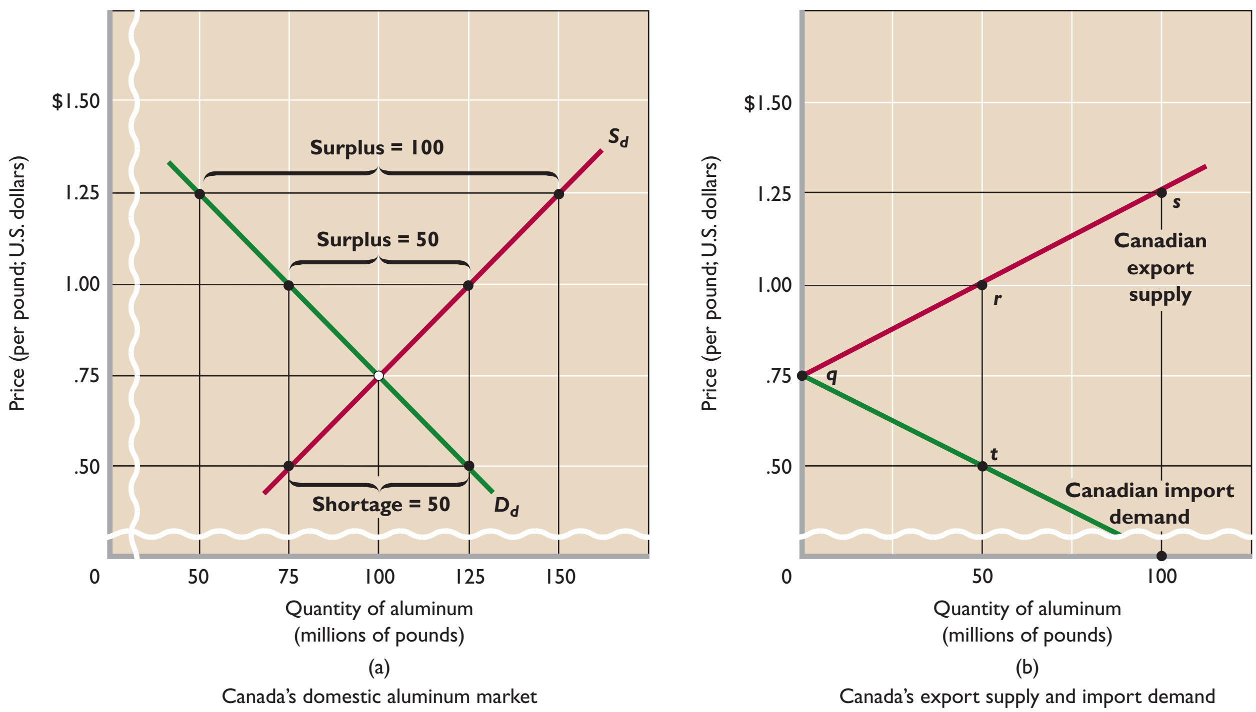
At an even lower world price, \$0.50, U.S. producers will supply only 50 million pounds. Because U.S. consumers want to buy 150 million pounds at that price, there is a domestic shortage of 100 million pounds. Imports will flow to the United States to make up the difference. That is, at a \$0.50 world price U.S. firms will supply 50 million pounds and 100 million pounds will be imported.

In Figure 38.3b we plot the U.S. **import demand curve** from these data. This *downsloping curve* shows the amounts of aluminum that will be imported at world prices below the \$1 U.S. domestic price. The relationship between world prices and imported amounts is inverse or negative. At a world price of \$1, domestic output will satisfy U.S. demand; imports will be zero (point *a*). But at \$0.75 the United States will import 50 million pounds of aluminum (point *x*); at \$0.50, the United States will import 100 million pounds (point *y*). Connecting points *a*, *x*, and *y* yields the *downsloping* U.S. import demand curve. It reveals that as world prices fall relative to U.S. domestic prices, U.S. imports increase.

Supply and Demand in Canada

We repeat our analysis in Figure 38.4, this time from the viewpoint of Canada. (We have converted Canadian dollar prices to U.S. dollar prices via the exchange rate.) Note that the domestic supply curve S_d and the domestic demand curve D_d for aluminum in Canada yield a domestic price of \$0.75, which is \$0.25 lower than the \$1 U.S. domestic price.

FIGURE 38.4 Canadian export supply and import demand. (a) At world prices above the $0.75 domestic price, production in Canada exceeds domestic consumption. At world prices below $0.75, domestic shortages occur. (b) Surpluses result in exports, and shortages result in imports. The Canadian export supply curve and import demand curve depict the relationships between world prices and exports or imports.



The analysis proceeds exactly as above except that the domestic price is now the Canadian price. If the world price is $0.75, Canadians will neither export nor import aluminum (giving us point *q* in Figure 38.4b). At world prices above $0.75, Canadian firms will produce more aluminum than Canadian consumers will buy. Canadian firms will export the surplus. At a $1 world price, Figure 38.4b tells us that Canada will have and export a domestic surplus of 50 million pounds (yielding point *r*). At $1.25, it will have and will export a domestic surplus of 100 million pounds (point *s*). Connecting these points yields the upsloping Canadian export supply curve, which reflects the domestic surpluses (and hence the exports) that occur when the world price exceeds the $0.75 Canadian domestic price.

At world prices below $0.75, domestic shortages occur in Canada. At a $0.50 world price, Figure 38.4a shows that Canadian consumers want to buy 125 million pounds of aluminum but Canadian firms will produce only 75 million pounds. The shortage will bring 50 million pounds of imports to Canada (point *t* in Figure 38.4b). The Canadian import demand curve in that figure shows the Canadian imports that will occur at all world aluminum prices below the $0.75 Canadian domestic price.

Equilibrium World Price, Exports, and Imports

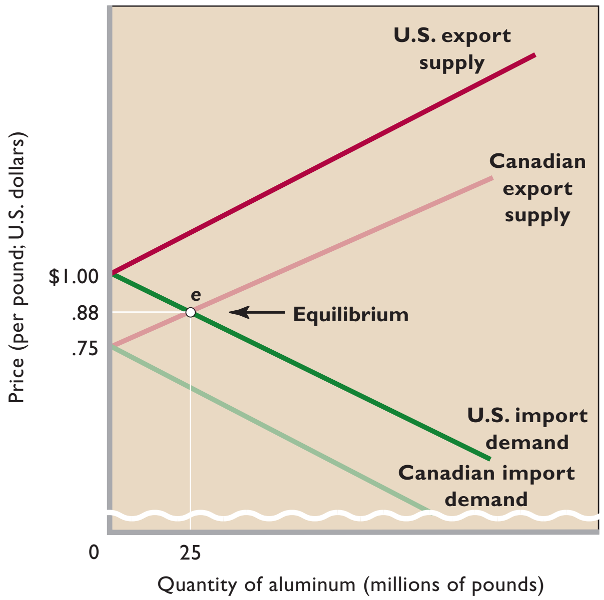
We now have the tools for determining the **equilibrium world price** of aluminum and the equilibrium world levels of exports and imports when the world is opened to trade. Figure 38.5 combines the U.S. export supply curve and import demand curve in Figure 38.3b and the Canadian export supply curve and import demand curve in Figure 38.4b. The two U.S. curves proceed rightward from the $1 U.S. domestic price; the two Canadian curves proceed rightward from the $0.75 Canadian domestic price.

International equilibrium occurs in this two-nation model where one nation's import demand curve intersects another nation's export supply curve. In this case the U.S. import demand curve intersects Canada's export supply curve at *e*. There, the world price of aluminum is $0.88. The Canadian export supply curve indicates that Canada will export 25 million pounds of aluminum at this price. Also at this price the United States will

WORKED PROBLEMS

W38.2
Equilibrium world price, exports, and imports

FIGURE 38.5 Equilibrium world price and quantity of exports and imports. In a two-nation world, the equilibrium world price (= $0.88) is determined by the intersection of one nation's export supply curve and the other nation's import demand curve. This intersection also decides the equilibrium volume of exports and imports. Here, Canada exports 25 million pounds of aluminum to the United States.



import 25 million pounds from Canada, indicated by the U.S. import demand curve. The $0.88 world price equates the quantity of imports demanded and the quantity of exports supplied (25 million pounds). Thus, there will be world trade of 25 million pounds of aluminum at $0.88 per pound.

Note that after trade, the single $0.88 world price will prevail in both Canada and the United States. Only one price for a standardized commodity can persist in a highly competitive world market. With trade, all consumers can buy a pound of aluminum for $0.88, and all producers can sell it for that price. This world price means that Canadians will pay more for aluminum with trade ($0.88) than without it ($0.75). The increased Canadian output caused by trade raises Canadian per-unit production costs and therefore raises the price of aluminum in Canada. The United States, however, pays less for aluminum with trade ($0.88) than without it ($1). The U.S. gain comes from Canada's comparative cost advantage in producing aluminum.

Why would Canada willingly send 25 million pounds of its aluminum output to the United States for U.S. consumption? After all, producing this output uses up scarce Canadian resources and drives up the price of aluminum for Canadians. Canadians are willing to export aluminum to the United States because Canadians gain the means—the U.S. dollars—to import other goods, say, computer software, from the United States. Canadian exports enable Canadians to acquire imports that have greater value to Canadians than the exported aluminum. Canadian exports to the United States finance Canadian imports from the United States.

QUICK REVIEW 38.4

- A nation will export a particular product if the world price exceeds the domestic price; it will import the product if the world price is less than the domestic price.
- In a two-country world model, equilibrium world prices and equilibrium quantities of exports and imports occur where one nation's export supply curve intersects the other nation's import demand curve.

Trade Barriers and Export Subsidies

LO38.4 Analyze the economic effects of tariffs and quotas.

While a nation as a whole gains from trade, trade may harm particular domestic industries and their workers. Those industries might seek to preserve their economic positions by persuading their respective governments to protect them from imports—perhaps through tariffs, import quotas, or other trade barriers.

Indeed, the public may be won over by the apparent plausibility ("Cut imports and prevent domestic unemployment") and the patriotic ring ("Buy American!") of the arguments. The alleged benefits of tariffs are immediate and clear-cut to the public, but the adverse effects cited by economists are obscure and dispersed over the entire economy. When political deal-making is added in—"You back tariffs for the apparel industry in my state, and I'll back tariffs for the auto industry in your state"—the outcome can be a politically robust network of trade barriers. These impediments to free international trade can take several forms.

Tariffs are excise taxes or "duties" on the dollar values or physical quantities of imported goods. They may be imposed to obtain revenue or to protect domestic firms. A **revenue tariff** is usually applied to a product that is not being produced domestically, for example, tin, coffee, or bananas in the case of the United States. Rates on revenue tariffs tend to be modest and are designed to provide the federal government with revenue. A **protective tariff** is

CONSIDER THIS . . .

Buy American?

Will "buying American" make Americans better off? No, says Dallas Federal Reserve economist W. Michael Cox:

> A common myth is that it is better for Americans to spend their money at home than abroad. The best way to expose the fallacy of this argument is to take it to its logical extreme. If it is better for me to spend my money here than abroad, then it is even better yet to buy in Texas than in New York, better yet to buy in Dallas than in Houston . . . in my own neighborhood . . . within my own family . . . to consume only what I can produce. Alone and poor.*

*"The Fruits of Free Trade," 2002 Annual Report, by W. Michael Cox and Richard Alm, p. 16, Federal Reserve Bank of Dallas. Used with permission.

implemented to shield domestic producers from foreign competition. These tariffs impede free trade by increasing the prices of imported goods and therefore shifting sales toward domestic producers. Although protective tariffs are usually not high enough to stop the importation of foreign goods, they put foreign producers at a competitive disadvantage. A tariff on imported auto tires, for example, would make domestically produced tires more attractive to consumers.

An **import quota** is a limit on the quantities or total values of specific items that are imported in some period. Once a quota is filled, further imports of that product are choked off. Import quotas are more effective than tariffs in impeding international trade. With a tariff, a product can go on being imported in large quantities. But with an import quota, all imports are prohibited once the quota is filled.

A **nontariff barrier (NTB)** includes onerous licensing requirements, unreasonable standards pertaining to product quality, or simply bureaucratic hurdles and delays in customs procedures. Some nations require that importers of foreign goods obtain licenses and then restrict the number of licenses issued. Although many nations carefully inspect imported agricultural products to prevent the introduction of potentially harmful insects, some countries use lengthy inspections to impede imports. Japan and the European countries frequently require that their domestic importers of foreign goods obtain licenses. By restricting the issuance of licenses, governments can limit imports.

A **voluntary export restriction (VER)** is a trade barrier by which foreign firms "voluntarily" limit the amount of their exports to a particular country. VERs have the same effect as import quotas and are agreed to by exporters to avoid more stringent tariffs or quotas. In the late 1990s, for example, Canadian producers of softwood lumber (fir, spruce, cedar, pine) agreed to a VER on exports to the United States under the threat of a permanently higher U.S. tariff.

ORIGIN OF THE IDEA

O38.2

Trade protectionism

An **export subsidy** consists of a government payment to a domestic producer of export goods and is designed to aid that producer. By reducing production costs, the subsidies enable the domestic firm to charge a lower price and thus to sell more exports in world markets. Two examples: Some European governments have heavily subsidized Airbus Industries, a European firm that produces commercial aircraft. The subsidies help Airbus compete against the American firm Boeing. The United States and other nations have subsidized domestic farmers to boost the domestic food supply. Such subsidies have artificially lowered export prices on agricultural produce.

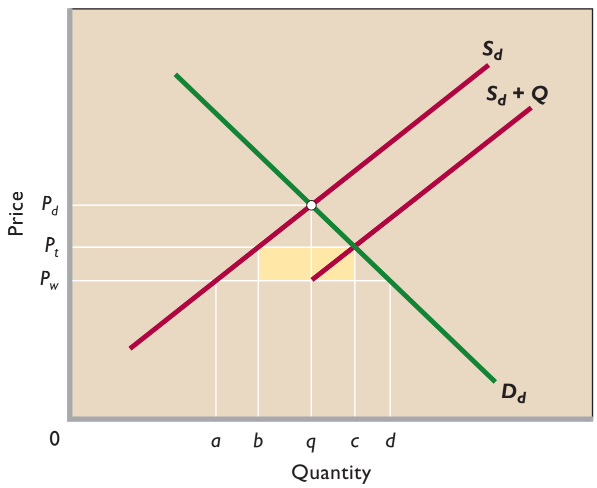
Later in this chapter we will discuss some of the specific arguments and appeals that are made to justify protection.

Economic Impact of Tariffs

We will confine our in-depth analysis of the effects of trade barriers to the two most common forms: tariffs and quotas. Once again, we turn to supply and demand analysis for help. Curves D_d and S_d in Figure 38.6 show domestic demand and supply for a product in which a nation, say, the United States, does *not* have a comparative advantage—for example, DVD players. (Disregard curve $S_d + Q$ for now.) Without world trade, the domestic price and output would be P_d and q, respectively.

Assume now that the domestic economy is opened to world trade and that Japan, which *does* have a comparative advantage in DVD players, begins to sell its players in the United States. We assume that with free trade the domestic price cannot differ from the world price, which here is P_w. At P_w domestic consumption is d and domestic production is a. The horizontal distance between the domestic supply and demand curves at P_w represents imports of ad.

FIGURE 38.6 The economic effects of a protective tariff or an import quota. A tariff that increases the price of a good from P_w to P_t will reduce domestic consumption from *d* to *c*. Domestic producers will be able to sell more output (*b* rather than *a*) at a higher price (P_t rather than P_w). Foreign exporters are injured because they sell less output (*bc* rather than *ad*). The yellow area indicates the amount of tariff paid by domestic consumers. An import quota of *bc* units has the same effect as the tariff, with one exception: The amount represented by the yellow area will go to foreign producers rather than to the domestic government.



Thus far, our analysis is similar to the analysis of world prices in Figure 38.3.

Direct Effects Suppose now that the United States imposes a tariff on each imported DVD player. The tariff, which raises the price of imported players from P_w to P_t, has four effects:

- ***Decline in consumption*** Consumption of DVD players in the United States declines from *d* to *c* as the higher price moves buyers up and to the left along their demand curve. The tariff prompts consumers to buy fewer players, and reallocate a portion of their expenditures to less desired substitute products. U.S. consumers are clearly injured by the tariff, since they pay $P_t - P_w$ more for each of the *c* units they buy at price P_t.
- ***Increased domestic production*** U.S. producers—who are not subject to the tariff—receive the higher price P_t per unit. Because this new price is higher than the pretariff world price P_w, the domestic DVD-player industry moves up and to the right along its supply curve S_d, increasing domestic output from *a* to *b*. Domestic producers thus enjoy both a higher price and expanded sales; this explains why domestic producers lobby for protective tariffs. But from a social point of view, the increase in domestic production from *a* to *b* means that the tariff permits domestic producers of players to bid resources away from other, more efficient, U.S. industries.
- ***Decline in imports*** Japanese producers are hurt. Although the sales price of each player is higher by $P_t - P_w$, that amount accrues to the U.S. government, not to Japanese producers. The after-tariff world price, or the per-unit revenue to Japanese producers, remains at P_w, but the volume of U.S. imports (Japanese exports) falls from *ad* to *bc*.
- ***Tariff revenue*** The yellow rectangle represents the amount of revenue the tariff yields. Total revenue from the tariff is determined by multiplying the tariff, $P_t - P_w$ per unit, by the number of players imported, *bc*. This tariff revenue is a transfer of income from consumers to government and does not represent any net change in the nation's economic well-being. The result is that government gains this portion of what consumers lose by paying more for DVD players.

Indirect Effect Tariffs have a subtle effect beyond what our supply and demand diagram can show. Because Japan sells fewer DVD players in the United States, it earns fewer dollars and so must buy fewer U.S. exports. U.S. export industries must then cut production and release resources. These are highly efficient industries, as we know from their comparative advantage and their ability to sell goods in world markets.

Tariffs directly promote the expansion of inefficient industries that do not have a comparative advantage; they also indirectly cause the contraction of relatively efficient industries that do have a comparative advantage. Put bluntly, tariffs cause resources to be shifted in the wrong direction—and that is not surprising. We know that specialization and world trade lead to more efficient use of world resources and greater world output. But protective tariffs reduce world trade. Therefore, tariffs also reduce efficiency and the world's real output.

Economic Impact of Quotas

We noted earlier that an import quota is a legal limit placed on the amount of some product that can be imported in a given year. Quotas have the same economic impact as a tariff, with one big difference: While tariffs generate revenue for the domestic government, a quota transfers that revenue to foreign producers.

Suppose in Figure 38.6 that, instead of imposing a tariff, the United States prohibits any imports of Japanese DVD players in excess of *bc* units. In other words, an import quota of *bc* players is imposed on Japan. We deliberately chose the size of this quota to be the same amount as imports would be under a $P_t - P_w$ tariff so that we can compare "equivalent" situations. As a consequence of the quota, the supply of players is $S_d + Q$ in the United States. This supply consists of the domestic supply plus the fixed amount *bc* (= Q) that importers will provide at each domestic price. The supply curve $S_w + Q$ does not extend below price P_w because Japanese producers would not export players to the United States at any price below P_w; instead, they would sell them to other countries at the world market price of P_w.

Most of the economic results are the same as those with a tariff. Prices of DVD players are higher (P_t instead of P_w) because imports have been reduced from *ad* to *bc*. Domestic consumption of DVD players is down from *d* to *c*. U.S. producers enjoy both a higher price (P_t rather than P_w) and increased sales (*b* rather than *a*).

The difference is that the price increase of $P_t - P_w$ paid by U.S. consumers on imports of *bc*—the yellow area—no longer goes to the U.S. Treasury as tariff (tax) revenue but flows to the Japanese firms that have acquired the quota rights to sell DVD players in the United States. For consumers in the United States, a tariff produces a better economic outcome than a quota, other things being the same. A tariff generates government revenue that can be used to cut other taxes or to finance public goods and services that benefit the United States. In contrast, the higher price created by quotas results in additional revenue for foreign producers.

QUICK REVIEW 38.5

- A tariff on a product increases its price, reduces its consumption, increases its domestic production, reduces its imports, and generates tariff revenue for the government; an import quota does the same, except a quota generates revenue for foreign producers rather than for the government imposing the quota.

Net Costs of Tariffs and Quotas

Figure 38.6 shows that tariffs and quotas impose costs on domestic consumers but provide gains to domestic producers and, in the case of tariffs, revenue to the federal government. The consumer costs of trade restrictions are calculated by determining the effect the restrictions have on consumer prices. Protection raises the price of a product in three ways: (1) The price of the imported product goes up; (2) the higher price of imports causes some consumers to shift their purchases to higher-priced domestically produced goods; and (3) the prices of domestically produced goods rise because import competition has declined.

Study after study finds that the costs to consumers substantially exceed the gains to producers and government. A sizable net cost or efficiency loss to society arises from trade protection. Furthermore, industries employ large amounts of economic resources to influence Congress to pass and retain protectionist laws. Because these rent-seeking efforts divert resources away from more socially desirable purposes, trade restrictions impose these additional costs on society as well.

Conclusion: The gains that U.S. trade barriers create for protected industries and their workers come at the expense of much greater losses for the entire economy. The result is economic inefficiency, reduced consumption, and lower standards of living.

The Case for Protection: A Critical Review

LO38.5 Analyze the validity of the most frequently presented arguments for protectionism.

Despite the logic of specialization and trade, there are still protectionists in some union halls, corporate boardrooms, and congressional conference rooms. What arguments do protectionists make to justify trade barriers? How valid are those arguments?

Military Self-Sufficiency Argument

The argument here is not economic but political-military: Protective tariffs are needed to preserve or strengthen industries that produce the materials essential for national defense. In an uncertain world, the political-military objectives (self-sufficiency) sometimes must take precedence over economic goals (efficiency in the use of world resources).

Unfortunately, it is difficult to measure and compare the benefit of increased national security against the cost of economic inefficiency when protective tariffs are imposed. The economist can only point out that when a nation levies tariffs to increase military self-sufficiency, it incurs economic costs.

All people in the United States would agree that relying on hostile nations for necessary military equipment is not a good idea, yet the self-sufficiency argument is open to serious abuse. Nearly every industry can claim that it makes direct or indirect contributions to national security and hence deserves protection from imports.

Diversification-for-Stability Argument

Highly specialized economies such as Saudi Arabia (based on oil) and Cuba (based on sugar) are dependent on international markets for their income. In these economies, wars, international political developments, recessions abroad, and random fluctuations in world supply and demand for one or two particular goods can cause deep declines in export revenues and therefore in domestic income. Tariff and quota protection are allegedly needed in such nations to enable greater industrial diversification. That way, these economies will not be so dependent on exporting one or two products to obtain the other goods they need. Such goods will be available domestically, thereby providing greater domestic stability.

There is some truth in this diversification-for-stability argument. But the argument has little or no relevance to the United States and other advanced economies. Also, the economic costs of diversification may be great; for example, one-crop economies may be highly inefficient at manufacturing.

Infant Industry Argument

The infant industry argument contends that protective tariffs are needed to allow new domestic industries to establish themselves. Temporarily shielding young domestic firms from the severe competition of more mature and more efficient foreign firms will give infant industries a chance to develop and become efficient producers.

This argument for protection rests on an alleged exception to the case for free trade. The exception is that young industries have not had, and if they face mature foreign competition will never have, the chance to make the long-run adjustments needed for larger scale and greater efficiency in production. In this view, tariff protection for such infant industries will correct a misallocation of world resources perpetuated by historically different levels of economic development between domestic and foreign industries.

There are some logical problems with the infant industry argument. In the developing nations it is difficult to determine which industries are the infants that are capable of achieving economic maturity and therefore deserving protection. Also, protective tariffs may persist even after industrial maturity has been realized.

Most economists feel that if infant industries are to be subsidized, there are better means than tariffs for doing so. Direct subsidies, for example, have the advantage of making explicit which industries are being aided and to what degree.

Protection-against-Dumping Argument

The protection-against-dumping argument contends that tariffs are needed to protect domestic firms from "dumping" by foreign producers. **Dumping** is the sale of a product in a foreign country at prices either below cost or below the prices commonly charged at home.

Economists cite two plausible reasons for this behavior. First, with regard to below-cost dumping, firms in country A may dump goods at below cost into country B in an attempt to drive their competitors in country B out of business. If the firms in country A succeed in driving their competitors in country B out of business, they will enjoy monopoly power and monopoly prices and profits on the goods they subsequently sell in country B. Their hope is that the longer-term monopoly profits will more than offset the losses from below-cost sales that must take place while they are attempting to drive their competitors in country B out of business.

Second, dumping that involves selling abroad at a price that is below the price commonly charged in the home country (but that is still at or above production costs) may be a form of price discrimination, which is charging different prices to different customers. As an example, a foreign seller that has a monopoly in its home market may find that it can maximize its overall profit by charging a high price in its monopolized domestic market while charging a lower price in the United States, where it must compete with U.S. producers. Curiously, it may pursue this strategy even if it makes no profit at all from its sales in the United States, where it must charge the competitive price. So why bother selling in the United States? Because the increase in overall production that comes about by exporting to the United States may allow the firm to obtain the per-unit cost savings often associated with large-scale production. These cost savings imply even higher profits in the monopolized domestic market.

Because dumping is an "unfair trade practice," most nations prohibit it. For example, where dumping is shown to injure U.S. firms, the federal government imposes tariffs called *antidumping duties* on the goods in question. But relatively few documented cases of dumping occur each year, and specific instances of unfair trade do not justify widespread, permanent tariffs. Moreover, antidumping duties can be abused. Often, what appears to be dumping is simply comparative advantage at work.

Increased Domestic Employment Argument

Arguing for a tariff to "save U.S. jobs" becomes fashionable when the economy encounters a recession (such as the severe recession of 2007–2009 in the United States).

In an economy that engages in international trade, exports involve spending on domestic output and imports reflect spending to obtain part of another nation's output. So, in this argument, reducing imports will divert spending on another nation's output to spending on domestic output. Thus, domestic output and employment will rise. But this argument has several shortcomings.

While imports may eliminate some U.S. jobs, they create others. Imports may have eliminated the jobs of some U.S. steel and textile workers in recent years, but other workers have gained jobs unloading ships, flying imported aircraft, and selling imported electronic equipment. Import restrictions alter the composition of employment, but they may have little or no effect on the volume of employment.

The *fallacy of composition*—the false idea that what is true for the part is necessarily true for the whole—is also present in this rationale for tariffs. All nations cannot simultaneously succeed in restricting imports while maintaining their exports; what is true for one nation is not true for all nations. The exports of one nation must be the imports of another nation. To the extent that one country is able to expand its economy through an excess of exports over imports, the resulting excess of imports over exports worsens another economy's unemployment problem. It is no wonder that tariffs and import quotas meant to achieve domestic full employment are called "beggar my neighbor" policies: They achieve short-run domestic goals by making trading partners poorer.

Moreover, nations adversely affected by tariffs and quotas are likely to retaliate, causing a "trade war" (more precisely, a *trade barrier war*) that will choke off trade and make all nations worse off. The **Smoot-Hawley Tariff Act** of 1930 is a classic example. Although that act was meant to reduce imports and stimulate U.S. production, the high tariffs it authorized prompted adversely affected nations to retaliate with tariffs equally high. International trade fell, lowering the output and income of all nations. Economic historians generally agree that the Smoot-Hawley Tariff Act was a contributing cause of the Great Depression.

Finally, forcing an excess of exports over imports cannot succeed in raising domestic employment over the long run. It is through U.S. imports that foreign nations earn dollars for buying U.S. exports. In the long run, a nation must import in order to export. The long-run impact of tariffs is not an increase in domestic employment but, at best, a reallocation of workers away from export industries and to protected domestic industries. This shift implies a less efficient allocation of resources.

Cheap Foreign Labor Argument

The cheap foreign labor argument says that domestic firms and workers must be shielded from the ruinous competition of countries where wages are low. If protection is not provided, cheap imports will flood U.S. markets and the prices of U.S. goods—along with the wages of U.S. workers—will be pulled down. That is, domestic living standards in the United States will be reduced.

This argument can be rebutted at several levels. The logic of the argument suggests that it is not mutually beneficial for rich and poor persons to trade with one another. However, that is not the case. A low-income farmworker may pick lettuce or tomatoes for a rich landowner, and both may benefit from the transaction. And both U.S. consumers and Chinese workers gain when they "trade" a pair of athletic shoes priced at $30 as opposed to U.S. consumers being restricted to buying a similar shoe made in the United States for $60.

Also, recall that gains from trade are based on comparative advantage, not on absolute advantage. Look back at Figure 38.1, where we supposed that the United States and Mexico had labor forces of exactly the same size. Noting the positions of the production possibilities curves, observe that U.S. labor can produce more of either good. Thus, it is more productive. Because of this greater productivity, we can expect wages and living standards to be higher for U.S. labor. Mexico's less productive labor will receive lower wages.

The cheap foreign labor argument suggests that, to maintain its standard of living, the United States should not trade with low-wage Mexico. What if it does not trade with Mexico? Will wages and living standards rise in the United States as a result? No. To obtain vegetables, the United States will have to reallocate a portion of its labor from its relatively more-efficient beef industry to its relatively less-efficient vegetables industry. As a result, the average productivity of U.S. labor will fall, as will real wages and living standards. The labor forces of both countries will have diminished standards of living because without specialization and trade they will have less output available to them. Compare column 4 with column 1 in Table 38.1 or points A' and Z' with A and Z in Figure 38.2 to confirm this point.

Another problem with the cheap foreign labor argument is that its proponents incorrectly focus on labor costs per hour when what really matters is labor costs per unit of output. As an example, suppose that a U.S. factory pays its workers $20 per hour while a factory in a developing country pays its workers $4 per hour. The proponents of the cheap foreign labor argument look at these numbers and conclude—incorrectly—that it is impossible for the U.S. factory to compete with the factory in the developing country. But this conclusion fails to take into account two crucial facts:

- What actually matters is labor costs per *unit of output*, not labor costs per *hour of work*.
- Differences in productivity typically mean that labor costs per unit of output are often nearly identical despite huge differences in hourly labor costs.

To see why these points matter so much, let's take into account how productive the two factories are. Because the U.S. factory uses much more sophisticated technology, better trained workers, and a lot more capital per worker, one worker in one hour can produce 20 units of output. Since the U.S. workers get paid $20 per hour, this means the U.S. factory's labor cost *per unit of output* is $1. The factory in the developing country is much less productive since it uses less efficient technology and its relatively untrained workers have a lot less machinery and equipment to work with. A worker there produces only 4 units per hour. Given the foreign wage of $4 per hour, this means that the labor cost per unit of output at the factory in the developing country is also $1.

As you can see, the lower wage rate per hour at the factory in the developing country does not translate into lower labor costs per unit—meaning that it won't be able to undersell its U.S. competitor just because its workers get paid lower wages per hour.

Proponents of the cheap foreign labor argument tend to focus exclusively on the large international differences that exist in labor costs per hour. They typically fail to mention that these differences in labor costs per hour are mostly the result of tremendously large differences in productivity and that these large differences in productivity serve to equalize labor costs per unit of output. As a result, firms in developing countries only *sometimes* have an advantage in terms of labor costs per unit of output. Whether they do in any specific situation will vary by industry and firm and will depend on differences in productivity as well as differences in labor costs per hour. For many goods, labor productivity in high-wage countries like the United States is so much higher than labor productivity in low-wage countries that it is actually cheaper *per unit of output* to manufacture those goods in high-wage countries. That is why, for instance, Intel still makes microchips in the United States and why most automobiles are still produced in the United States, Japan, and Europe rather than in low-wage countries.

QUICK REVIEW 38.6

- Most rationales for trade protections are special-interest requests that, if followed, would create gains for protected industries and their workers at the expense of greater losses for the economy.

Multilateral Trade Agreements and Free-Trade Zones

LO38.6 Identify and explain the objectives of GATT, WTO, EU, eurozone, and NAFTA, and discuss offshoring and trade adjustment assistance.

Aware of the detrimental effects of trade wars and the general weaknesses of arguments for trade protections, nations have worked to lower tariffs worldwide. Their pursuit of freer trade has been aided by recently emerged special-interest groups that have offset the more-established special-interest groups that have traditionally supported tariffs and quotas. Specifically, lower tariffs are now supported by exporters of goods and services, importers of foreign components used in "domestic" products, and domestic sellers of imported products.

General Agreement on Tariffs and Trade

In 1947, 23 nations, including the United States, signed the **General Agreement on Tariffs and Trade (GATT).** GATT was based on three principles: (1) equal, nondiscriminatory trade treatment for all member nations; (2) the reduction of tariffs by multilateral negotiation; and (3) the elimination of import quotas. Basically, GATT provided a forum for the multilateral negotiation of reduced trade barriers.

Since the Second World War, member nations have completed eight "rounds" of GATT negotiations to reduce trade barriers. The eighth round of negotiations began in Uruguay in 1986. After seven years of complex discussions, in 1993 a new agreement was reached by the 128 nations that were by that time members of GATT. The Uruguay Round agreement took effect on January 1, 1995, and its provisions were phased in through 2005.

Under this agreement, tariffs on thousands of products were eliminated or reduced, with overall tariffs dropping by 33 percent. The agreement also liberalized government rules that in the past impeded the global market for such services as advertising, legal services, tourist services, and financial services. Quotas on imported textiles and apparel were phased out and replaced with tariffs. Other provisions reduced agricultural subsidies paid to farmers and protected intellectual property (patents, trademarks, and copyrights) against piracy.

World Trade Organization

The Uruguay Round agreement established the **World Trade Organization (WTO)** as GATT's successor. Some 159 nations belonged to the WTO in 2013. The WTO oversees trade agreements reached by the member nations, and rules on trade disputes among them. It also provides forums for further rounds of trade negotiations. The ninth and latest round of negotiations—the **Doha Development Agenda**—was launched in Doha, Qatar, in late 2001. (The trade rounds occur over several years in several venues but are named after the city or country of origination.) The negotiations are aimed at further reducing tariffs and quotas, as well as agricultural subsidies that distort trade. You can get an update on the status of the complex negotiations at **www.wto.org**.

GATT and the WTO have been positive forces in the trend toward liberalized world trade. The trade rules agreed upon by the member nations provide a strong and necessary bulwark against the protectionism called for by the special-interest groups in the various nations.

For that reason and others, the WTO is quite controversial. Critics are concerned that rules crafted to expand international trade and investment enable firms to circumvent national laws that protect workers and the environment. Critics ask: What good are minimum-wage laws, worker-safety laws, collective-bargaining rights, and environmental laws if firms can easily shift their production to nations that have weaker laws or if consumers can buy goods produced in those countries?

Proponents of the WTO respond that labor and environmental protections should be pursued directly in nations that have low standards and via international organizations other than the WTO. These issues should not be linked to the process of trade liberalization, which confers widespread economic benefits across nations. Moreover, say proponents of the WTO, many environmental and labor concerns are greatly overblown. Most world trade is among advanced industrial countries, not between them and countries that have lower environmental and labor standards. Moreover, the free flow of goods and resources raises output and income in the developing nations. Historically, such increases in living standards have eventually resulted in stronger, not weaker, protections for the environment and for workers.

The European Union

Countries have also sought to reduce tariffs by creating regional free-trade zones. The most dramatic example is the **European Union (EU).** Initiated in 1958 as the Common Market, in 2003 the EU comprised 15 European nations—Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, and the United Kingdom. In 2004, the EU expanded by 10 additional European countries—Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia. The 2007 addition of Bulgaria and Romania plus the 2013 addition of Croatia expanded the EU to 28 nations.

The EU has abolished tariffs and import quotas on nearly all products traded among the participating nations and established a common system of tariffs applicable to all goods received from nations outside the EU. It has also liberalized the movement of capital and labor within the EU and has created common policies in other economic matters of joint concern, such as agriculture, transportation, and business practices.

EU integration has achieved for Europe what the U.S. constitutional prohibition on tariffs by individual states has achieved for the United States: increased regional specialization, greater productivity, greater output, and faster economic growth. The free flow of goods and services has created large markets for EU industries. The resulting economies of large-scale production have enabled these industries to achieve much lower costs than they could have achieved in their small, single-nation markets.

One of the most significant accomplishments of the EU was the establishment of the so-called **eurozone** or euro area in the early 2000s. As of 2013, 17 members of the EU (Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovenia, Slovakia, and Spain) use the euro as a common currency. Notably, the United Kingdom, Denmark, and Sweden have opted not to use the common currency, at least for now. But gone are French francs, German marks, Italian liras, and other national currencies that were once used by eurozone countries.

Economists expect the adoption of the euro to raise the standard of living in the eurozone nations over time. By ending the inconvenience and expense of exchanging currencies, the euro has enhanced the free flow of goods, services, and resources among the eurozone members. Companies that previously sold products in only one or two European nations have found it easier to price and sell their products in all 17 eurozone countries. The euro has also allowed consumers and businesses to comparison shop for outputs and inputs, and this capability has increased competition, reduced prices, and lowered costs.

North American Free Trade Agreement

In 1993 Canada, Mexico, and the United States created a major free-trade zone. The **North American Free Trade Agreement (NAFTA)** established a free-trade area that has about the same combined output as the EU but encompasses a much larger geographic area. NAFTA has eliminated tariffs and other trade barriers between Canada, Mexico, and the United States for most goods and services.

Critics of NAFTA feared that it would cause a massive loss of U.S. jobs as firms moved to Mexico to take advantage of lower wages and weaker regulations on pollution and workplace safety. Also, they were concerned that Japan and South Korea would build plants in Mexico and transport goods tariff-free to the United States, further hurting U.S. firms and workers.

In retrospect, critics were much too pessimistic. Since the passage of NAFTA in 1993, employment in the United States has increased by more than 20 million workers. NAFTA has increased trade among Canada, Mexico, and the United States and has enhanced the standard of living in all three countries.

QUICK REVIEW 38.7

- The General Agreement on Tariffs and Trade (GATT) of 1947 reduced tariffs and quotas and established a process for numerous subsequent rounds of multinational trade negotiations that have liberalized international trade.
- The World Trade Organization (WTO)—GATT's successor—rules on trade disputes and provides forums for negotiations on further rounds of trade liberalization. The current round of negotiations is called the Doha Development Agenda.
- The European Union (EU) and the North American Free Trade Agreement (NAFTA) have reduced internal trade barriers among their member nations by establishing multination free-trade zones.

Recognizing Those Hurt by Free Trade

Shifts in patterns of comparative advantage and removal of long-standing trade protection can hurt specific groups of workers. For example, the erosion of the United States' once strong comparative advantage in steel has caused production plant shutdowns and layoffs in the U.S. steel industry. The textile and apparel industries in the United States face similar difficulties. Clearly, not everyone wins from free trade (or freer trade). Some workers lose.

Trade Adjustment Assistance

The **Trade Adjustment Assistance Act** of 2002 introduced some innovative policies to help those hurt by shifts in international trade patterns. The law provides cash assistance (beyond unemployment insurance) for up to 78 weeks for workers displaced by imports or plant relocations abroad. To obtain the assistance, workers must participate in job searches, training programs, or remedial education. Also provided are relocation allowances to help displaced workers move geographically to new jobs within the United States. Refundable tax credits for health insurance serve as payments to help workers maintain their insurance coverage during the retraining and job-search period. Workers who are 50 years of age or older are eligible for "wage insurance," which replaces some of the difference in pay (if any) between their old and new jobs. Many economists support trade adjustment assistance because it not only helps workers hurt by international trade but also helps create the political support necessary to reduce trade barriers and export subsidies.

But not all economists favor trade adjustment assistance. Loss of jobs from imports, sending some work abroad, and plant relocations to other countries are only a small fraction (about 4 percent in recent years) of total job losses in the economy each year. Many workers also lose their jobs because of changing patterns of demand, changing technology, bad management, and other dynamic aspects of a market economy. Some critics ask, "What makes losing one's job to international trade worthy of such special treatment, compared to losing one's job to, say, technological change or domestic competition?" Economists can find no totally satisfying answer.

Offshoring of Jobs

Not only are some U.S. jobs lost because of international trade, but some are lost because of globalization of resource markets. In recent years U.S. firms have found the outsourcing of work abroad to be increasingly profitable. Economists call this business activity **offshoring**—shifting work previously done by American workers to workers located in other nations. Offshoring is not a new practice but traditionally has involved components for U.S. manufacturing goods. For example, Boeing has long offshored the production of major airplane parts for its "American" aircraft.

Recent advances in computer and communications technology have enabled U.S. firms to offshore service jobs such as data entry, book composition, software coding, call-center operations, medical transcription, and claims processing to countries such as India. Where offshoring

LAST WORD

Petition of the Candlemakers, 1845

French Economist Frédéric Bastiat (1801–1850) Devastated the Proponents of Protectionism by Satirically Extending Their Reasoning to Its Logical and Absurd Conclusions.

Petition of the Manufacturers of Candles, Waxlights, Lamps, Candlesticks, Street Lamps, Snuffers, Extinguishers, and of the Producers of Oil Tallow, Rosin, Alcohol, and, Generally, of Everything Connected with Lighting.

TO MESSIEURS THE MEMBERS OF THE CHAMBER OF DEPUTIES.

Gentlemen—You are on the right road. You reject abstract theories, and have little consideration for cheapness and plenty. Your chief care is the interest of the producer. You desire to emancipate him from external competition, and reserve the national market for national industry.

We are about to offer you an admirable opportunity of applying your—what shall we call it? your theory? No; nothing is more deceptive than theory; your doctrine? your system? your principle? but you dislike doctrines, you abhor systems, and as for principles, you deny that there are any in social economy: we shall say, then, your practice, your practice without theory and without principle.

We are suffering from the intolerable competition of a foreign rival, placed, it would seem, in a condition so far superior to ours for the production of light, that he absolutely inundates our national market with it at a price fabulously reduced. The moment he shows himself, our trade leaves us—all consumers apply to him; and a branch of native industry, having countless ramifications, is all at once rendered completely stagnant. This rival . . . is no other than the Sun.

What we pray for is, that it may please you to pass a law ordering the shutting up of all windows, skylights, dormer windows, outside and inside shutters, curtains, blinds, bull's-eyes; in a word, of all openings, holes, chinks, clefts, and fissures, by or through which the light of the sun has been in use to enter houses, to the prejudice of the meritorious manufacturers with which we flatter ourselves we have accommodated our country,—a country which, in gratitude, ought not to abandon us now to a strife so unequal.

If you shut up as much as possible all access to natural light, and create a demand for artificial light, which of our French manufacturers will not be encouraged by it? If more tallow is consumed, then there must be more oxen and sheep; and, consequently, we shall behold the multiplication of artificial meadows, meat, wool, hides, and, above all, manure, which is the basis and foundation of all agricultural wealth.

The same remark applies to navigation. Thousands of vessels will proceed to the whale fishery; and, in a short time, we shall possess a navy capable of maintaining the honor of France, and gratifying the patriotic aspirations of your petitioners, the undersigned candle-makers and others.

Only have the goodness to reflect, Gentlemen, and you will be convinced that there is, perhaps, no Frenchman, from the wealthy coalmaster to the humblest vender of lucifer matches, whose lot will not be ameliorated by the success of this our petition.

Source: Frédéric Bastiat, *Economic Sophisms* (Irvington-on-Hudson, NY: The Foundation for Economic Education, Inc., 1996), abridged. Used with permission of Foundation for Economic Education, **www.FEE.org**.

occurs, some of the value added in the production process accrues to foreign countries rather than the United States. So part of the income generated from the production of U.S. goods is paid to foreigners, not to American workers.

Offshoring is a wrenching experience for many Americans who lose their jobs, but it is not necessarily bad for the overall economy. Offshoring simply reflects growing specialization and international trade in services, or, more descriptively, "tasks." That growth has been made possible by recent trade agreements and new information and communication technologies. As with trade in goods, trade in services reflects comparative advantage and is beneficial to both trading parties. Moreover, the United States has a sizable trade surplus with other nations in services. The United States gains by specializing in high-valued services such as transportation services, accounting services, legal services, and advertising services, where it still has a comparative advantage. It then "trades" to obtain lower-valued services such as call-center and data-entry work, for which comparative advantage has gone abroad.

Offshoring also increases the demand for complementary jobs in the United States. Jobs that are close substitutes for existing U.S. jobs are lost, but complementary jobs in the United States are expanded. For example, the lower price of writing software code in India may mean a lower cost of software sold in the United States and abroad. That, in turn, may create more jobs for U.S.-based workers such as software designers, marketers, and distributors. Moreover, offshoring may encourage domestic investment and the expansion of firms in the United States by reducing their production costs and keeping them competitive worldwide. In some instances, "offshoring jobs" may equate to "importing competitiveness." Entire firms that might otherwise disappear abroad may remain profitable in the United States only because they can offshore some of their work.

QUICK REVIEW 38.8

- Increased international trade and offshoring of jobs have harmed some specific U.S. workers and have led to policies such as trade adjustment assistance to try to help them with their transitions to new lines of work.

SUMMARY

LO38.1 List and discuss several key facts about international trade.

The United States leads the world in the combined volume of exports and imports. Other major trading nations are Germany, Japan, the western European nations, and the Asian economies of China, South Korea, Taiwan, and Singapore. The United States' principal exports include chemicals, agricultural products, consumer durables, semiconductors, and aircraft; principal imports include petroleum, automobiles, metals, household appliances, and computers.

LO38.2 Define comparative advantage, and demonstrate how specialization and trade add to a nation's output.

World trade is based on three considerations: the uneven distribution of economic resources among nations, the fact that efficient production of various goods requires particular techniques or combinations of resources, and the differentiated products produced among nations.

Mutually advantageous specialization and trade are possible between any two nations if they have different domestic opportunity-cost ratios for any two products. By specializing on the basis of comparative advantage, nations can obtain larger real incomes with fixed amounts of resources. The terms of trade determine how this increase in world output is shared by the trading nations. Increasing (rather than constant) opportunity costs limit specialization and trade.

LO38.3 Describe how differences between world prices and domestic prices prompt exports and imports.

A nation's export supply curve shows the quantities of a product the nation will export at world prices that exceed the domestic price (the price in a closed, no-international-trade economy). A nation's import demand curve reveals the quantities of a product it will import at world prices below the domestic price.

In a two-nation model, the equilibrium world price and the equilibrium quantities of exports and imports occur where one nation's export supply curve intersects the other nation's import demand curve. A nation will export a particular product if the world price exceeds the domestic price; it will import the product if the world price is less than the domestic price. The country with the lower costs of production will be the exporter and the country with the higher costs of production will be the importer.

LO38.4 Analyze the economic effects of tariffs and quotas.

Trade barriers take the form of protective tariffs, quotas, nontariff barriers, and "voluntary" export restrictions. Export subsidies also distort international trade. Supply and demand analysis demonstrates that protective tariffs and quotas increase the prices and reduce the quantities demanded of the affected goods. Sales by foreign exporters diminish; domestic producers, however, gain higher prices and enlarged sales. Consumer losses from trade restrictions greatly exceed producer and government gains, creating an efficiency loss to society.

LO38.5 Analyze the validity of the most frequently presented arguments for protectionism.

The strongest arguments for protection are the infant industry and military self-sufficiency arguments. Most other arguments for protection are interest-group appeals or reasoning fallacies that emphasize producer interests over consumer interests or stress the immediate effects of trade barriers while ignoring long-run consequences.

The cheap foreign labor argument for protection fails because it focuses on labor costs per hour rather than on what really matters, labor costs per unit of output. Due to higher productivity, firms in high-wage countries like the United States can have lower wage costs per unit of output than competitors in low-wage countries. Whether they do will depend on how their particular wage and productivity levels compare with those of their competitors in low-wage countries.

LO38.6 Identify and explain the objectives of GATT, WTO, EU, eurozone, and NAFTA, and discuss offshoring and trade adjustment assistance.

In 1947 the General Agreement on Tariffs and Trade (GATT) was formed to encourage nondiscriminatory treatment for all member nations, to reduce tariffs, and to eliminate import quotas. The Uruguay Round of GATT negotiations (1993) reduced tariffs and quotas, liberalized trade in services, reduced agricultural subsidies, reduced pirating of intellectual property, and phased out quotas on textiles.

GATT's successor, the World Trade Organization (WTO), had 159 member nations in 2013. It implements WTO agreements, rules on trade disputes between members, and provides forums for continued discussions on trade liberalization. The latest round of trade negotiations—the Doha Development Agenda—began in late 2001 and as of 2013 was still in progress.

Free-trade zones liberalize trade within regions. Two examples of free-trade arrangements are the 28-member European Union (EU) and the North American Free Trade Agreement (NAFTA), comprising Canada, Mexico, and the United States. Seventeen EU nations have abandoned their national currencies for a common currency called the euro.

The Trade Adjustment Assistance Act of 2002 recognizes that trade liberalization and increased international trade can create job loss for many workers. The Act therefore provides cash assistance, education and training benefits, health care subsidies, and wage subsidies (for persons aged 50 or older) to qualified workers displaced by imports or relocations of plants from the United States to abroad.

Offshoring is the practice of shifting work previously done by Americans in the United States to workers located in other nations. Although offshoring reduces some U.S. jobs, it lowers production costs, expands sales, and therefore may create other U.S. jobs. Less than 4 percent of all job losses in the United States each year are caused by imports, offshoring, and plant relocation abroad.

TERMS AND CONCEPTS

labor-intensive goods
land-intensive goods
capital-intensive goods
opportunity-cost ratio
comparative advantage
principle of comparative advantage
terms of trade
trading possibilities line
gains from trade
world price
domestic price
export supply curve
import demand curve
equilibrium world price
tariffs
revenue tariff
protective tariff
import quota
nontariff barrier (NTB)
voluntary export restriction (VER)
export subsidy
dumping
Smoot-Hawley Tariff Act
General Agreement on Tariffs and Trade (GATT)
World Trade Organization (WTO)
Doha Development Agenda
European Union (EU)
eurozone
North American Free Trade Agreement (NAFTA)
Trade Adjustment Assistance Act
offshoring

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Quantitatively, how important is international trade to the United States relative to the importance of trade to other nations? What country is the United States' most important trading partner, quantitatively? With what country does the United States have the largest trade deficit? **LO38.1**
2. Distinguish among land-, labor-, and capital-intensive goods, citing an example of each without resorting to book examples. How do these distinctions relate to international trade? How do distinctive products, unrelated to resource intensity, relate to international trade? **LO38.1, LO38.2**
3. Explain: "The United States can make certain toys with greater productive efficiency than can China. Yet we import those toys from China." Relate your answer to the ideas of Adam Smith and David Ricardo. **LO38.2**

4. Suppose Big Country can produce 80 units of X by using all its resources to produce X or 60 units of Y by devoting all its resources to Y. Comparable figures for Small Nation are 60 units of X and 60 units of Y. Assuming constant costs, in which product should each nation specialize? Explain why. What are the limits of the terms of trade between these two countries? How would rising costs (rather than constant costs) affect the extent of specialization and trade between these two countries? **LO38.2**
5. What is an export supply curve? What is an import demand curve? How do such curves relate to the determination of the equilibrium world price of a tradable good? **LO38.3**
6. Why is a quota more detrimental to an economy than a tariff that results in the same level of imports as the quota? What is the net outcome of either tariffs or quotas for the world economy? **LO38.4**
7. "The potentially valid arguments for tariff protection—military self-sufficiency, infant industry protection, and diversification for stability—are also the most easily abused." Why are these arguments susceptible to abuse? **LO38.4**
8. Evaluate the effectiveness of artificial trade barriers, such as tariffs and import quotas, as a way to achieve and maintain full employment throughout the U.S. economy. How might such policies reduce unemployment in one U.S. industry but increase it in another U.S. industry? **LO38.4**
9. In 2012, manufacturing workers in the United States earned average compensation of $35.67 per hour. That same year, manufacturing workers in Mexico earned average compensation of $6.36 per hour. How can U.S. manufacturers possibly compete? Why isn't all manufacturing done in Mexico and other low-wage countries? **LO38.4**
10. How might protective tariffs reduce both the imports and the exports of the nation that levies tariffs? In what way do foreign firms that "dump" their products onto the U.S. market in effect provide bargains to American consumers? How might the import competition lead to quality improvements and cost reductions by American firms? **LO38.4**
11. Identify and state the significance of each of the following trade-related entities: (*a*) the WTO; (*b*) the EU; (*c*) the eurozone; and (*d*) NAFTA. **LO38.6**
12. What form does trade adjustment assistance take in the United States? How does such assistance promote political support for free-trade agreements? Do you think workers who lose their jobs because of changes in trade laws deserve special treatment relative to workers who lose their jobs because of other changes in the economy, say, changes in patterns of government spending? **LO38.6**
13. What is offshoring of white-collar service jobs and how does that practice relate to international trade? Why has offshoring increased over the past few decades? Give an example (other than that in the textbook) of how offshoring can eliminate some American jobs while creating other American jobs. **LO38.6**
14. **LAST WORD** What was the central point that Bastiat was trying to make in his imaginary petition of the candlemakers?

REVIEW QUESTIONS

1. In Country A, a worker can make 5 bicycles per hour. In Country B, a worker can make 7 bicycles per hour. Which country has an absolute advantage in making bicycles? **LO38.2**
   a. Country A.
   b. Country B.
2. In Country A, the production of 1 bicycle requires using resources that could otherwise be used to produce 11 lamps. In Country B, the production of 1 bicycle requires using resources that could otherwise be used to produce 15 lamps. Which country has a comparative advantage in making bicycles? **LO38.2**
   a. Country A.
   b. Country B.
3. True or False: If Country B has an absolute advantage over Country A in producing bicycles, it will also have a comparative advantage over Country A in producing bicycles. **LO38.2**
4. Suppose that the opportunity-cost ratio for sugar and almonds is $4S \equiv 1A$ in Hawaii but $1S \equiv 2A$ in California. Which state has the comparative advantage in producing almonds? **LO38.2**
   a. Hawaii.
   b. California.
   c. Neither.
5. Suppose that the opportunity-cost ratio for fish and lumber is $1F \equiv 1L$ in Canada but $2F \equiv 1L$ in Iceland. Then ____________ should specialize in producing fish while ____________ should specialize in producing lumber. **LO38.2**
   a. Canada; Iceland.
   b. Iceland; Canada.
6. Suppose that the opportunity-cost ratio for watches and cheese is $1C \equiv 1W$ in Switzerland but $1C \equiv 4W$ in Japan. At which of the following international exchange ratios (terms of trade) will Switzerland and Japan be willing to specialize and engage in trade with each other. **LO38.2**
   *Select **one or more** answers from the choices shown.*
   a. $1C \equiv 3W$.
   b. $1C \equiv \frac{1}{2}W$.
   c. $1C \equiv 5W$.
   d. $\frac{1}{2}C \equiv 1W$.
   e. $2C \equiv 1W$.
7. We see quite a bit of international trade in the real world. And trade is driven by specialization. So why don't we see full specialization—for instance, all cars in the world being made in South Korea, or all the mobile phones in the world

being made in China? Choose the best answer from among the following choices. **LO38.2**

a. High tariffs.
b. Extensive import quotas.
c. Increasing opportunity costs.
d. Increasing returns.

8. Which of the following are benefits of international trade? **LO38.2**

*Choose **one or more** answers from the choices shown.*

a. A more efficient allocation of resources.
b. A higher level of material well-being.
c. Gains from specialization.
d. Promoting competition.
e. Deterring monopoly.
f. Reducing the threat of war.

9. True or False: If a country is open to international trade, the domestic price can differ from the international price. **LO38.3**

10. Suppose that the current international price of wheat is \$6 per bushel and that the United States is currently exporting 30 million bushels per year. If the United States suddenly became a closed economy with respect to wheat, would the domestic price of wheat in the United States end up higher or lower than \$6? **LO38.3**

a. Higher.
b. Lower.
c. The same.

11. Suppose that if Iceland and Japan were both closed economies, the domestic price of fish would be \$100 per ton in Iceland and \$90 per ton in Japan. If the two countries decided to open up to international trade with each other, which of the following could be the equilibrium international price of fish once they begin trading? **LO38.3**

a. \$75.
b. \$85.
c. \$95.
d. \$105.

12. Draw a domestic supply-and-demand diagram for a product in which the United States does not have a comparative advantage. What impact do foreign imports have on domestic price and quantity? On your diagram show a protective tariff that eliminates approximately one-half of the assumed imports. What are the price-quantity effects of this tariff on (*a*) domestic consumers, (*b*) domestic producers, and (*c*) foreign exporters? How would the effects of a quota that creates the same amount of imports differ? **LO38.4**

13. American apparel makers complain to Congress about competition from China. Congress decides to impose either a tariff or a quota on apparel imports from China. Which policy would Chinese apparel manufacturers prefer? **LO38.4**

a. Tariff.
b. Quota.

PROBLEMS

1. Assume that the comparative-cost ratios of two products—baby formula and tuna fish—are as follows in the nations of Canswicki and Tunata:

Canswicki: 1 can baby formula ≡ 2 cans tuna fish
Tunata: 1 can baby formula ≡ 4 cans tuna fish

In what product should each nation specialize? Which of the following terms of trade would be acceptable to both nations: (*a*) 1 can baby formula ≡ $2\frac{1}{2}$ cans tuna fish; (*b*) 1 can baby formula ≡ 1 can tuna fish; (*c*) 1 can baby formula ≡ 5 cans tuna fish? **LO38.2**

2. The accompanying hypothetical production possibilities tables are for New Zealand and Spain. Each country can produce apples and plums. Plot the production possibilities data for each of the two countries separately. Referring to your graphs, answer the following: **LO38.2**

New Zealand's Production Possibilities Table (Millions of Bushels)

	Production Alternatives			
Product	**A**	**B**	**C**	**D**
Apples	0	20	40	60
Plums	15	10	5	0

Spain's Production Possibilities Table (Millions of Bushels)

	Production Alternatives			
Product	**R**	**S**	**T**	**U**
Apples	0	20	40	60
Plums	60	40	20	0

a. What is each country's cost ratio of producing plums and apples?
b. Which nation should specialize in which product?
c. Show the trading possibilities lines for each nation if the actual terms of trade are 1 plum for 2 apples. (Plot these lines on your graph.)
d. Suppose the optimum product mixes before specialization and trade were alternative B in New Zealand and alternative S in Spain. What would be the gains from specialization and trade?

3. The following hypothetical production possibilities tables are for China and the United States. Assume that before specialization and trade the optimal product mix for China is alternative B and for the United States is alternative U. **LO38.2**

a. Are comparative-cost conditions such that the two areas should specialize? If so, what product should each produce?

b. What is the total gain in apparel and chemical output that would result from such specialization?

c. What are the limits of the terms of trade? Suppose that the actual terms of trade are 1 unit of apparel for $1\frac{1}{2}$ units of chemicals and that 4 units of apparel are exchanged for 6 units of chemicals. What are the gains from specialization and trade for each nation?

	China Production Possibilities					
Product	**A**	**B**	**C**	**D**	**E**	**F**
Apparel (in thousands)	30	24	18	12	6	0
Chemicals (in tons)	0	6	12	18	24	30

	U.S. Production Possibilities					
Product	**R**	**S**	**T**	**U**	**V**	**W**
Apparel (in thousands)	10	8	6	4	2	0
Chemicals (in tons)	0	4	8	12	16	20

4. Refer to Figure 3.6, page 63. Assume that the graph depicts the U.S. domestic market for corn. How many bushels of corn, if any, will the United States export or import at a world price of $1, $2, $3, $4, and $5? Use this information to construct the U.S. export supply curve and import demand curve for corn. Suppose that the only other corn-producing nation is France, where the domestic price is $4. Which country will export corn; which county will import it? **LO38.3**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

CHAPTER **39**

The Balance of Payments, Exchange Rates, and Trade Deficits

Learning Objectives

LO39.1 Explain how currencies of different nations are exchanged when international transactions take place.

LO39.2 Analyze the balance sheet the United States uses to account for the international payments it makes and receives.

LO39.3 Discuss how exchange rates are determined in currency markets that have flexible exchange rates.

LO39.4 Describe the difference between flexible exchange rates and fixed exchange rates.

LO39.5 Explain the current system of managed floating exchange rates.

LO39.6 Identify the causes and consequences of recent U.S. trade deficits.

If you take a U.S. dollar to the bank and ask to exchange it for U.S. currency, you will get a puzzled look. If you persist, you may get a dollar's worth of change: One U.S. dollar can buy exactly one U.S. dollar. But on April 21, 2013, for example, 1 U.S. dollar could buy 1,837 Colombian pesos, 0.97 Australian dollar, 0.66 British pound, 1.03 Canadian dollars, 0.77 European euro, 99.55 Japanese yen, or 12.29 Mexican pesos. What explains this seemingly haphazard array of exchange rates?

In Chapter 38 we examined comparative advantage as the underlying economic basis of world trade and discussed the effects of barriers to free trade. Now we introduce the highly important monetary and financial aspects of international trade.

International Financial Transactions

LO39.1 Explain how currencies of different nations are exchanged when international transactions take place.

This chapter focuses on international financial transactions, the vast majority of which fall into two broad categories: international trade and international asset transactions. International trade involves either purchasing or selling currently produced goods or services across an international border. Examples include an Egyptian firm exporting cotton to the United States and an American company hiring an Indian call center to answer its phones. International asset transactions involve the transfer of the property rights to either real or financial assets between the citizens of one country and the citizens of another country. It includes activities like buying foreign stocks or selling your house to a foreigner.

These two categories of international financial transactions reflect the fact that whether they are from different countries or the same country, individuals and firms can only exchange two things with each other: currently produced goods and services or preexisting assets. With regard to assets, however, money is by far the most commonly exchanged asset. Only rarely would you ever find a barter situation in which people directly exchanged other assets—such as trading a car for 500 shares of Microsoft stock or a cow for 30 chickens and a tank of diesel fuel.

As a result, there are two basic types of transactions:

- People trading either goods or services for money.
- People trading assets for money.

In either case, money flows from the buyers of the goods, services, or assets to the sellers of the goods, services, or assets.

When the people engaged in any such transactions are both from places that use the same currency, what type of money to use is not an issue. Americans from California and Wisconsin will use their common currency, the dollar. People from France and Germany will use their common currency, the euro. However, when the people involved in an exchange are from places that use different currencies, intermediate asset transactions have to take place: the buyers must convert their own currencies into the currencies that the sellers use and accept.

As an example, consider the case of an English software design company that wants to buy a supercomputer made by an American company. The American company sells these high-powered machines for $300,000. To pay for the machine, the English company has to convert some of the money it has (British pounds sterling) into the money that the American company will accept (U.S. dollars). This process is not difficult. As we will soon explain in detail, there are many easy-to-use foreign exchange markets in which those who wish to sell pounds and buy dollars can interact with others who wish to sell dollars and buy pounds. The demand and supply created by these two groups determine the equilibrium exchange rate, which, in turn, determines how many pounds our English company will have to convert to pay for the supercomputer. For instance, if the exchange rate is $2 = £1, then the English company will have to convert £150,000 to obtain the $300,000 necessary to purchase the computer.

QUICK REVIEW 39.1

- International financial transactions involve trade either in currently produced goods and services or in preexisting assets.
- Exports of goods, services, and assets create inflows of money, while imports cause outflows of money.
- If buyers and sellers use different currencies, then foreign exchange transactions take place so that the exporter can be paid in his or her own currency.

The Balance of Payments

LO39.2 Analyze the balance sheet the United States uses to account for the international payments it makes and receives.

A nation's **balance of payments** is the sum of all the financial transactions that take place between its residents and the residents of foreign nations. Most of these transactions fall into the two main categories that we have just discussed: international trade and international asset

TABLE 39.1 The U.S. Balance of Payments, 2012 (in Billions)

CURRENT ACCOUNT		
(1) U.S. goods exports	$+1,564	
(2) U.S. goods imports	−2,299	
(3) *Balance on goods*		$−735
(4) U.S. exports of services	+630	
(5) U.S. imports of services	−435	
(6) *Balance on services*		+195
(7) *Balance on goods and services*		−540
(8) Net investment income	+199*	
(9) Net transfers	−134	
(10) **Balance on current account**		**−475**
CAPITAL AND FINANCIAL ACCOUNT		
Capital account		
(11) *Balance on capital account*		+6
Financial account		
(12) Foreign purchases of assets in the United States	+418†	
(13) U.S. purchases of assets abroad	+51†	
(14) *Balance on financial account*		+469
(15) **Balance on capital and financial account**		**+475**
		$ 0

*Includes other, less significant, categories of income.
†Includes one-half of a $66 billion statistical discrepancy that is listed in the capital account.

Source: U.S. Department of Commerce, Bureau of Economic Analysis, **www.bea.gov**. Preliminary 2012 data. The export and import data are on a "balance-of-payment basis," and usually vary from the data on exports and imports reported in the National Income and Product Accounts.

transactions. As a result, nearly all the items included in the balance of payments are things such as exports and imports of goods, exports and imports of services, and international purchases and sales of financial and real assets. But the balance of payments also includes international transactions that fall outside of these main categories—things such as tourist expenditures, interest and dividends received or paid abroad, debt forgiveness, and remittances made by immigrants to their relatives back home.

The U.S. Commerce Department's Bureau of Economic Analysis compiles a balance-of-payments statement each year. This statement summarizes all of the millions of payments that individuals and firms in the United States receive from foreigners as well as all of the millions of payments that individuals and firms in the United States make to foreigners. It shows "flows" of inpayments of money *to* the United States and outpayments of money *from* the United States. For convenience, all of these money payments are given in terms of dollars. This is true despite the fact that some of them actually may have been made using foreign currencies—as when, for instance, an American company converts dollars into euros to buy something from an Italian company. When including this outpayment of money from the United States, the accountants who compile the balance-of-payments statement use the number of dollars the American company converted—rather than the number of euros that were actually used to make the purchase.

Table 39.1 is a simplified balance-of-payments statement for the United States in 2012. Because most international financial transactions fall into only two categories—international trade and international asset exchanges—the balance-of-payments statement is organized into two broad categories. *The current account* located at the top of the table primarily treats international trade. *The capital and financial account* at the bottom of the table primarily treats international asset exchanges.

Current Account

The top portion of Table 39.1 that mainly summarizes U.S. trade in currently produced goods and services is called the **current account.** Items 1 and 2 show U.S. exports and imports of goods (merchandise) in 2012. U.S. exports have a *plus* (+) sign because they are a *credit;* they generate flows of money toward the United States. U.S. imports have a *minus* (−) sign because they are a *debit;* they cause flows of money out of the United States.

Balance on Goods Items 1 and 2 in Table 39.1 reveal that in 2012 U.S. goods exports of $1,564 billion were less than U.S. goods imports of $2,299 billion. A country's *balance of trade on goods* is the difference between its exports and its imports of goods. If exports exceed imports, the result is a surplus on the balance of goods. If imports exceed exports, there is a trade deficit on the balance of goods. We note in item 3 that in 2012 the United States incurred a trade deficit on goods of $735 billion.

Balance on Services The United States exports not only goods, such as airplanes and computer software, but also services, such as insurance, consulting, travel, and investment advice, to residents of foreign nations. Item 4 in Table 39.1 shows that these service "exports" totaled $630 billion in 2012. Since they generate flows of money toward the United States, they are a credit (thus the + sign). Item 5 indicates that the United States "imports" similar services from foreigners. Those service imports were $435 billion in 2012, and since they generate flows of money out of the United States, they are a debit (thus the − sign). Summed together, items 4 and 5 indicate that the balance on services (item 6) in 2012 was $195 billion. The **balance on goods and services** shown as item 7 is the difference between U.S. exports of goods and services (items 1 and 4) and U.S. imports of goods and services (items 2 and 5). In 2012, U.S. imports of goods and services exceeded U.S. exports of goods and services by $540 billion. So a **trade deficit** of that amount occurred. In contrast, a **trade surplus** occurs when exports of goods and services exceed imports of goods and services. (Global Perspective 39.1 shows U.S. trade deficits and surpluses with selected nations.)

Balance on Current Account Items 8 and 9 are not items relating directly to international trade in goods and services. But they are listed as part of the current account (which is mostly about international trade in goods and services) because they are international financial flows that in some sense compensate for things that can be conceptualized as being *like* international trade in either goods or services. For instance, item 8, *net investment income*, represents the difference between (1) the interest and dividend payments foreigners paid U.S. citizens and companies for the services provided by U.S. capital invested abroad ("exported" capital) and (2) the interest and dividends the U.S. citizens and companies paid for the services provided by foreign capital invested here ("imported" capital). Observe that in 2012 U.S. net investment income was a positive $199 billion.

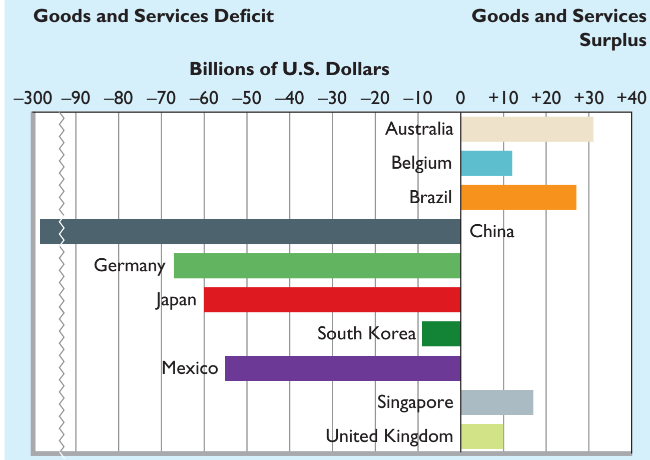
Item 9 shows net transfers, both public and private, between the United States and the rest of the world. Included here is foreign aid, pensions paid to U.S. citizens living abroad, and remittances by immigrants to relatives abroad. These $134 billion of transfers are net U.S. outpayments (and therefore listed as a negative number in Table 39.1). They are listed as part of the current account because they can be thought of as the financial flows that accompany the exporting of goodwill and the importing of "thank you notes."

By adding all transactions in the current account, we obtain the **balance on current account** shown in item 10. In 2012 the United States had a current account deficit of $475 billion. This means that the U.S. current account transactions created outpayments from the United States greater than inpayments to the United States.

GLOBAL PERSPECTIVE 39.1

U.S. Trade Balances in Goods and Services, Selected Nations, 2012

The United States has large trade deficits in goods and services with several nations, in particular, China, Germany, and Japan.



Source: Bureau of Economic Analysis, **www.bea.gov**.

Capital and Financial Account

The bottom portion of the current account statement summarizes U.S. international asset transactions. It is called the **capital and financial account** and consists of two separate accounts: the *capital account* and the *financial account*.

Capital Account The capital account mainly measures debt forgiveness—which is an asset transaction because the person forgiving a debt essentially hands the IOU back to the borrower. It is a "net" account (one that can be either + or −). The +$6 billion listed in line 11 tells us that in 2012 foreigners forgave $6 billion more of debt owed to them by Americans than Americans forgave debt owed to them by foreigners. The + sign indicates a credit; it is an "on-paper" inpayment (asset transfer) by the net amount of debt forgiven.

Financial Account The financial account summarizes international asset transactions having to do with international purchases and sales of real or financial assets. Line 12 lists the amount of foreign purchases of assets in the United States. It has a + sign because any purchase of an American-owned asset by a foreigner generates a flow of money toward the American who sells the asset. Line 13 lists U.S. purchases of assets abroad. These normally have a − sign because such purchases generate a flow of money from the Americans who buy foreign assets toward the foreigners who sell them those assets. But the value in line 13 is positive because in 2012 American sales to foreigners of assets located abroad exceeded American purchases from foreigners of assets located abroad. That excess generated a net flow of $51 billion from foreigners to Americans.

Items 12 and 13 combined yielded a $469 billion balance on the financial account for 2012 (line 14). In 2012 the United States "exported" $418 billion of ownership of its real and financial assets and "imported" $51 billion. Thought of differently, this surplus in the financial account brought in income of $469 billion to the United States. The **balance on the capital and financial account** (line 15) is $475 billion. It is the sum of the $6 billion credit on the capital account and the $469 billion surplus on the financial account. Observe that this $475 billion surplus in the capital and financial account equals the $475 billion deficit in the current account. This is not an accident. The two numbers always equal—or "balance." That's why the statement is called the *balance* of payments. It has to balance. Let's see why.

Why the Balance?

The balance on the current account and the balance on the capital and financial account must always sum to zero because any deficit or surplus in the current account automatically creates an offsetting entry in the capital and financial account. People can only trade one of two things with each other: currently produced goods and services or preexisting assets. Therefore, if trading partners have an imbalance in their trade of currently produced goods and services, the only way to make up for that imbalance is with a net transfer of assets from one party to the other.

To see why this is true, suppose that John (an American) makes shoes and Henri (a Swiss citizen) makes watches and that the pair only trade with each other. Assume that their financial assets consist entirely of money, with each beginning the year with $1,000 in his bank account. Suppose that this year John exports $300 of shoes to Henri and imports $500 of watches from Henri. John therefore ends the year with a $200 goods deficit with Henri.

John and Henri's goods transactions, however, also result in asset exchanges that cause a net transfer of assets from John to Henri equal in size to John's $200 goods deficit with Henri. This is true because Henri pays John $300 for his shoes while John pays Henri $500 for his watches. The *net* result of these opposite-direction asset movements is that $200 of John's initial assets of $1,000 are transferred to Henri. This is unavoidable because the $300 John receives from his exports pays for only the first $300 of his $500 of imports. The only way for John to pay for the remaining $200 of imports is for him to transfer $200 of his initial asset holdings to Henri. Consequently, John's assets decline by $200 from $1,000 to $800, and Henri's assets rise from $1,000 to $1,200.

Consider how the transaction between John and Henri affects the U.S. balance-of-payments statement (Table 39.1), other things equal. John's $200 goods deficit with Henri shows up in the U.S. current account as a −$200 entry in the balance on goods account (line 3) and carries down to a −$200 entry in the balance on current account (line 10).

In the capital and financial account, this $200 is recorded as +$200 in the account labeled foreign purchases of assets in the United States (line 12). This +$200 then carries down to the balance on capital and financial account (line 15). Think of it this way: Henri has in essence used $200 worth of watches to purchase $200 of John's initial $1,000 holding of assets. The +$200 entry in line 12 (foreign purchases of assets in the United States) simply recognizes this fact. This +$200 exactly offsets the −$200 in the current account.

Thus, the balance of payments always balances. Any current account deficit or surplus in the top half of the statement automatically generates an offsetting international asset transfer that shows up in the capital and financial account in the bottom half of the statement. More specifically, current account deficits simultaneously generate transfers of assets to foreigners, while current account surpluses automatically generate transfers of assets from foreigners.

Official Reserves, Payments Deficits, and Payments Surpluses

Some of the foreign purchases of assets in the United States (line 12, Table 39.1) and the U.S. assets purchased abroad (line 13) are of so-called official reserves. **Official reserves** consist of foreign currencies, certain reserves held with the International Monetary Fund, and stocks of gold. These reserves are owned by governments or their central banks. For simplicity, we will assume for now that the entire stock of U.S. official reserves consists of foreign currency so we can speak of official reserves and foreign currency reserves interchangeably.

Although the balance of payments must always sum to zero, as in Table 39.1, in some years a net sale of official reserves by a nation's treasury or central bank occurs in the process of bringing the capital and financial account into balance with the current account. In such years, a **balance-of-payments deficit** is said to occur. This deficit is in a subset of the overall balance statement and *is not a deficit in the overall account.* Remember, the overall balance of payments is always in balance. But in this case the balancing of the overall account includes sales of official reserves to create an inflow of dollars to the United States. In selling foreign currency in the foreign exchange market, the treasury or central bank must draw down its stock of reserves. This drawdown is an indicator of a balance-of-payments deficit.

These net sales of official reserves in the foreign exchange market show up as a *plus* (+) item on the U.S. balance-of-payments statement, specifically as foreign purchases of U.S. assets (line 12). They are a credit or an inflow of dollars to the United States, just as are John's proceeds from the sale of $200 of his assets to Henri in our previous example.

In other years, the capital and financial account balances the current account because of government purchases of official reserves from foreigners. The treasury or central bank engineers this balance by selling dollars to obtain foreign currency, and then adding the newly acquired foreign currency to its stock of official reserves. In these years, a **balance-of-payments surplus** is said to exist. This payments surplus therefore can be thought of as either net purchases of official reserves in the balance of payments or, alternatively, as the resulting increase in the stock of official reserves held by the government.

Net purchases of official reserves by the treasury or central bank appear on the U.S. balance sheet as U.S. purchases of foreign assets (line 13)—a *negative* (−) item. These purchases are a debit because they represent an outflow of dollars.

A balance-of-payments deficit is not necessarily bad, just as a balance-of-payments surplus is not necessarily good. Both simply happen. However, any nation's official reserves are limited. Persistent payments deficits must be financed by drawing down those reserves, which would ultimately deplete the reserves. That nation would have to adopt policies to correct its balance of payments. Such policies might require painful macroeconomic adjustments, trade barriers and similar restrictions, or a major depreciation of its currency. For this reason, nations strive for payments balance, at least over several-year periods.

WORKED PROBLEMS

W39.1
Balance of payments

QUICK REVIEW 39.2

- A nation's balance-of-payments statement summarizes all of the international financial transactions that take place between its residents and the residents of all foreign nations. It includes the current account balance and the capital and financial account balance.
- The current account balance is a nation's exports of goods and services less its imports of goods and services plus its net investment income and net transfers.
- The capital and financial account balance includes the net amount of the nation's debt forgiveness as well as the nation's sale of real and financial assets to people living abroad less its purchases of real and financial assets from foreigners.
- The current account balance and the capital and financial account balance always sum to zero because any current account imbalance automatically generates an offsetting international asset transfer.
- A balance-of-payments deficit exists when net sales of official reserves (mainly foreign currency) occur in the balance-of-payments statement; in contrast, a balance-of-payments surplus exists when net purchases of official reserves occur.

Flexible Exchange Rates

LO39.3 Discuss how exchange rates are determined in currency markets that have flexible exchange rates.

Both the size and the persistence of a nation's balance-of-payments deficits and surpluses and the adjustments it must make to correct those imbalances depend on the system of exchange rates being used. There are two pure types of exchange-rate systems:

- A **flexible- or floating-exchange-rate system** through which demand and supply determine exchange rates and in which no government intervention occurs.

KEY GRAPH

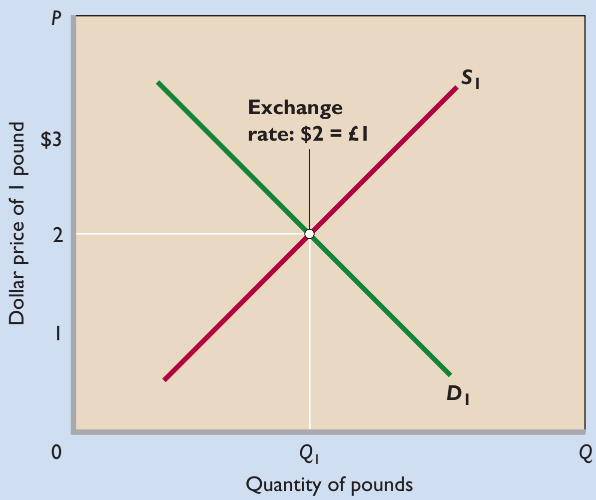


FIGURE 39.1 The market for foreign currency (pounds). The intersection of the demand-for-pounds curve D_1 and the supply-of-pounds curve S_1 determines the equilibrium dollar price of pounds, here, \$2. That means that the exchange rate is \$2 = £1. Not shown, an increase in demand for pounds or a decrease in supply of pounds will increase the dollar price of pounds and thus cause the pound to appreciate. Also not shown, a decrease in demand for pounds or an increase in the supply of pounds will reduce the dollar price of pounds, meaning that the pound has depreciated.

QUICK QUIZ FOR FIGURE 39.1

1. Which of the following statements is true?
   a. The quantity of pounds demanded falls when the dollar appreciates.
   b. The quantity of pounds supplied declines as the dollar price of the pound rises.
   c. At the equilibrium exchange rate, the pound price of \$1 is £$\frac{1}{2}$.
   d. The dollar appreciates if the demand for pounds increases.
2. At the price of \$2 for £1 in this figure:
   a. the dollar-pound exchange rate is unstable.
   b. the quantity of pounds supplied equals the quantity demanded.
   c. the dollar price of £1 equals the pound price of \$1.
   d. U.S. goods exports to Britain must equal U.S. goods imports from Britain.
3. Other things equal, a leftward shift of the demand curve in this figure:
   a. would depreciate the dollar.
   b. would create a shortage of pounds at the previous price of \$2 for £1.
   c. might be caused by a major recession in the United States.
   d. might be caused by a significant rise of real interest rates in Britain.
4. Other things equal, a rightward shift of the supply curve in this figure would:
   a. depreciate the dollar and might be caused by a significant rise of real interest rates in Britain.
   b. depreciate the dollar and might be caused by a significant fall of real interest rates in Britain.
   c. appreciate the dollar and might be caused by a significant rise of real interest rates in the United States.
   d. appreciate the dollar and might be caused by a significant fall of real interest rates in the United States.

Answers: 1. c; 2. b; 3. c; 4. c

- A **fixed-exchange-rate system** through which governments determine exchange rates and make necessary adjustments in their economies to maintain those rates.

We begin by looking at flexible exchange rates. Let's examine the rate, or price, at which U.S. dollars might be exchanged for British pounds. In **Figure 39.1 (Key Graph)** we show demand D_1 and supply S_1 of pounds in the currency market.

The *demand-for-pounds curve* is downsloping because all British goods and services will be cheaper to the United States if pounds become less expensive to the United States. That is, at lower dollar prices for pounds, the United States can obtain more pounds and therefore more British goods and services per dollar. To buy those cheaper British goods, U.S. consumers will increase the quantity of pounds they demand.

The *supply-of-pounds curve* is upsloping because the British will purchase more U.S. goods when the dollar price of pounds rises (that is, as the pound price of dollars falls). When the British buy more U.S. goods, they supply a greater quantity of pounds to the foreign exchange market. In other words, they must exchange pounds for dollars to purchase U.S. goods. So, when the dollar price of pounds rises, the quantity of pounds supplied goes up.

The intersection of the supply curve and the demand curve will determine the dollar price of pounds. Here, that price (exchange rate) is \$2 for £1. At this exchange rate, the quantities of pounds supplied and demanded are equal; neither a shortage nor a surplus of pounds occurs.

Depreciation and Appreciation

An exchange rate determined by market forces can, and often does, change daily like stock and bond prices. When the dollar price of pounds *rises*, for example, from \$2 = £1 to \$3 = £1, the dollar has *depreciated* relative to the pound (and the pound has appreciated relative to the dollar). When a currency depreciates, more units of it (dollars) are needed to buy a single unit of some other currency (a pound).

When the dollar price of pounds *falls*, for example, from \$2 = £1 to \$1 = £1, the dollar has *appreciated* relative to the pound. When a currency appreciates, fewer units of it (dollars) are needed to buy a single unit of some other currency (pounds).

In our U.S.-Britain illustrations, depreciation of the dollar means an appreciation of the pound, and vice versa. When the dollar price of a pound jumps from \$2 = £1 to \$3 = £1, the pound has appreciated relative to the dollar because it takes fewer pounds to buy \$1. At \$2 = £1, it took £$\frac{1}{2}$ to buy \$1; at \$3 = £1, it takes only £$\frac{1}{3}$ to buy \$1. Conversely, when the dollar appreciated relative to the pound, the pound depreciated relative to the dollar. More pounds were needed to buy a dollar.

In general, the relevant terminology and relationships between the U.S. dollar and another currency are as follows.

- Dollar price of foreign currency increases ≡ dollar depreciates relative to the foreign currency ≡ foreign currency price of dollar decreases ≡ foreign currency appreciates relative to the dollar.
- Dollar price of foreign currency decreases ≡ dollar appreciates relative to the foreign currency ≡ foreign currency price of dollar increases ≡ foreign currency depreciates relative to the dollar.

Determinants of Exchange Rates

What factors would cause a nation's currency to appreciate or depreciate in the market for foreign exchange? Here are three generalizations:

- If the demand for a nation's currency increases (other things equal), that currency will appreciate; if the demand declines, that currency will depreciate.
- If the supply of a nation's currency increases, that currency will depreciate; if the supply decreases, that currency will appreciate.
- If a nation's currency appreciates, some foreign currency depreciates relative to it.

With these generalizations in mind, let's examine the determinants of exchange rates—the factors that shift the demand or supply curve for a certain currency. As we do so, keep in mind that the other-things-equal assumption is always in force. Also note that we are discussing factors *that change the exchange rate*, not things that change *as a result of* a change in the exchange rate.

Changes in Tastes Any change in consumer tastes or preferences for the products of a foreign country may alter the demand for that nation's currency and change its exchange rate. If technological advances in U.S. wireless phones make them more attractive to British consumers and businesses, then the British will supply more pounds in the exchange market to purchase more U.S. wireless phones. The supply-of-pounds curve will shift to the right, causing the pound to depreciate and the dollar to appreciate.

In contrast, the U.S. demand-for-pounds curve will shift to the right if British woolen apparel becomes more fashionable in the United States. So the pound will appreciate and the dollar will depreciate.

Relative Income Changes A nation's currency is likely to depreciate if its growth of national income is more rapid than that of other countries. Here's why: A country's imports vary directly with its income level. As total income rises in the United States, people there buy both more domestic goods and more foreign goods. If the U.S. economy is expanding rapidly and the British economy is stagnant, U.S. imports of British goods, and therefore U.S. demands for pounds, will increase. The dollar price of pounds will rise, so the dollar will depreciate.

Relative Inflation Rate Changes Other things equal, changes in the relative rates of inflation of two nations change their relative price levels and alter the exchange rate between their currencies. The currency of the nation with the higher inflation rate—the more rapidly rising price level—tends to depreciate. Suppose, for example, that inflation is zero percent in Great Britain and 5 percent in the United States so that prices, on average, are rising by 5 percent per year in the United States while, on average, remaining unchanged in Great Britain. U.S. consumers will seek out more of the now relatively lower-priced British goods, increasing the demand for pounds. British consumers will purchase less of the now relatively higher-priced U.S. goods, reducing the supply of pounds. This combination of increased demand for pounds and reduced supply of pounds will cause the pound to appreciate and the dollar to depreciate.

According to the **purchasing-power-parity theory,** exchange rates should eventually adjust such that they equate the purchasing power of various currencies. If a certain market basket of identical products costs \$10,000 in the United States and £5,000 in Great Britain, the exchange rate should move to \$2 = £1. That way, a dollar spent in the United States will buy exactly as much output as it would if it were first converted to pounds (at the \$2 = £1 exchange rate) and used to buy output in Great Britain.

In terms of our example, 5 percent inflation in the United States will increase the price of the market basket from \$10,000 to \$10,500, while the zero percent inflation in Great Britain will leave the market basket priced at £5,000. For purchasing power parity to hold, the exchange rate would have to move from \$2 = £1 to \$2.10 = £1. That means the dollar therefore would depreciate and the pound would appreciate. In practice, however, not all exchange rates move precisely to equate the purchasing power of various currencies and thereby achieve "purchasing power parity," even over long periods.

Relative Interest Rates Changes in relative interest rates between two countries may alter their exchange rate. Suppose that real interest rates rise in the United States but stay constant in Great Britain. British citizens will then find the United States a more attractive place in which to loan money directly or loan money indirectly by buying bonds. To make these loans, they will have to supply pounds in the foreign exchange market to obtain dollars. The increase in the supply of pounds results in depreciation of the pound and appreciation of the dollar.

Changes in Relative Expected Returns on Stocks, Real Estate, and Production Facilities International investing extends beyond buying foreign bonds. It includes international investments in stocks and real estate as well as foreign purchases of factories and production facilities. Other things equal, the extent of this foreign investment depends on relative expected returns. To make the investments, investors in one country must sell their currencies to purchase the foreign currencies needed for the foreign investments.

For instance, suppose that investing in England suddenly becomes more popular due to a more positive outlook regarding expected returns on stocks, real estate, and production facilities there. U.S. investors therefore will sell U.S. assets to buy more assets in England. The U.S. assets will be sold for dollars, which will then be brought to the foreign exchange market and exchanged for pounds, which will in turn be used to purchase British assets. The increased demand for pounds in the foreign exchange market will cause the pound to appreciate and therefore the dollar to depreciate relative to the pound.

Speculation Currency speculators are people who buy and sell currencies with an eye toward reselling or repurchasing them at a profit. Suppose speculators expect the U.S. economy to (1) grow more rapidly than the British economy and (2) experience more rapid inflation than Britain. These expectations translate into an anticipation that the pound will appreciate and the dollar will depreciate. Speculators who are holding dollars will therefore try to convert them into pounds. This effort will increase the demand for pounds and cause the dollar price of pounds to rise (that is, cause the dollar to depreciate). A self-fulfilling prophecy occurs: The pound appreciates and the dollar depreciates because speculators act on the belief that these changes will in fact take place. In this way, speculation can cause changes in exchange rates. (We discuss currency speculation in more detail in this chapter's Last Word.)

Table 39.2 has more illustrations of the determinants of exchange rates; the table is worth careful study.

QUICK REVIEW 39.3

- In a system in which exchange rates are flexible (free to float), exchange rates are determined by the demand for and supply of individual national currencies in the foreign exchange market.
- Determinants of flexible exchange rates (factors that shift currency supply and demand curves) include (a) changes in tastes; (b) relative national incomes; (c) relative inflation rates; (d) real interest rates; (e) relative expected returns on stocks, real estate, and production facilities; and (f) speculation.

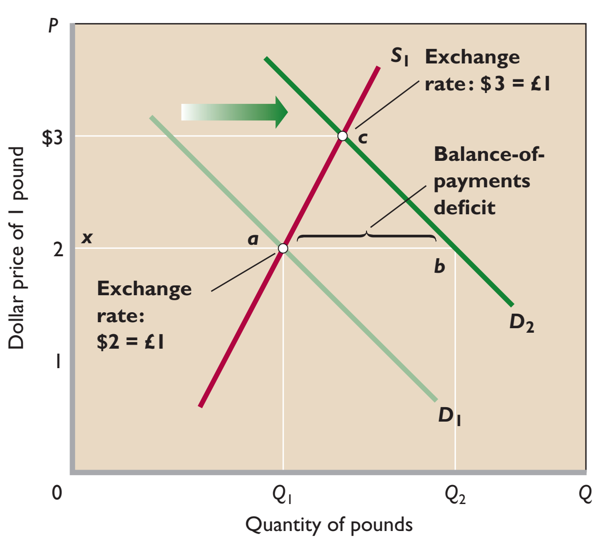
Flexible Rates and the Balance of Payments

Flexible exchange rates have an important feature: They automatically adjust and eventually eliminate balance-of-payments deficits or surpluses. We can explain this idea through Figure 39.2, in which S_1 and D_1 are the supply and demand curves for pounds from Figure 39.1. The equilibrium exchange rate of \$2 = £1 means that there is no balance-of-payments deficit or surplus between the United States and Britain. At that exchange rate, the quantity of pounds demanded by U.S. consumers to import British goods, buy British transportation and insurance services, and pay interest and dividends on British investments in the United States equals the amount of pounds supplied by the British in buying U.S. exports, purchasing

TABLE 39.2 Determinants of Exchange Rates: Factors That Change the Demand for or the Supply of a Particular Currency and Thus Alter the Exchange Rate

Determinant	Examples
Change in tastes	Japanese electronic equipment declines in popularity in the United States (Japanese yen depreciates; U.S. dollar appreciates). European tourists reduce visits to the United States (U.S. dollar depreciates; European euro appreciates).
Change in relative incomes	England encounters a recession, reducing its imports, while U.S. real output and real income surge, increasing U.S. imports (British pound appreciates; U.S. dollar depreciates).
Change in relative inflation rates	Switzerland experiences a 3% inflation rate compared to Canada's 10% rate (Swiss franc appreciates; Canadian dollar depreciates).
Change in relative real interest rates	The Federal Reserve drives up interest rates in the United States, while the Bank of England takes no such action (U.S. dollar appreciates; British pound depreciates).
Changes in relative expected returns on stocks, real estate, or production facilities	Corporate tax cuts in the United States raise expected after-tax investment returns in the United States relative to those in Europe (U.S. dollar appreciates; the euro depreciates).
Speculation	Currency traders believe South Korea will have much greater inflation than Taiwan (South Korean won depreciates; Taiwanese dollar appreciates). Currency traders think Norway's interest rates will plummet relative to Denmark's rates (Norway's krone depreciates; Denmark's krone appreciates).

FIGURE 39.2 Adjustments under flexible exchange rates and fixed exchange rates. Under flexible exchange rates, a shift in the demand for pounds from D_1 to D_2, other things equal, would cause a U.S. balance-of-payments deficit *ab*. That deficit would be corrected by a change in the exchange rate from \$2 = £1 to \$3 = £1. Under fixed exchange rates, the United States would cover the shortage of pounds *ab* by selling official reserves (here pounds), restricting trade, implementing exchange controls, or enacting a contractionary stabilization policy.



services from the United States, and making interest and dividend payments on U.S. investments in Britain. The United States would have no need to either draw down or build up its official reserves to balance its payments.

Suppose tastes change and U.S. consumers buy more British automobiles; the U.S. inflation rate increases relative to Britain's; or interest rates fall in the United States compared to those in Britain. Any or all of these changes will increase the U.S. demand for British pounds, for example, from D_1 to D_2 in Figure 39.2.

If the exchange rate remains at the initial \$2 = £1, a U.S. balance-of-payments deficit will occur in the amount of *ab*. At the \$2 = £1 rate, U.S. consumers will demand the quantity of pounds shown by point *b*, but Britain will supply only the amount shown by *a*. There will be a shortage of pounds. But this shortage will not last because this is a competitive foreign exchange market. Instead, the dollar price of pounds will rise (the dollar will depreciate) until the balance-of-payments deficit is eliminated. That occurs at the new equilibrium exchange rate of \$3 = £1, where the quantities of pounds demanded and supplied are again equal.

To explain why the increase in the dollar price of pounds—the dollar depreciation—eliminates the balance-of-payments deficit *ab* in Figure 39.2, we need to reemphasize that the exchange rate links all domestic (U.S.) prices with all foreign (British) prices. The dollar price of a foreign good is found by multiplying the foreign price by the exchange rate (in dollars per unit of the foreign currency). At an exchange rate of \$2 = £1, a British automobile priced at £15,000 will cost a U.S. consumer \$30,000 (= 15,000 × \$2).

A change in the exchange rate alters the prices of all British goods to U.S. consumers and all U.S. goods to British buyers. The shift in the exchange rate (here from \$2 = £1 to \$3 = £1) changes the relative attractiveness of U.S. imports and exports and restores equilibrium in the U.S. (and British) balance of payments. From the U.S. view, as the dollar price of pounds changes from \$2 to \$3, the British auto priced at £15,000, which formerly cost a U.S. consumer $30,000, now costs $45,000 (= 15,000 × $3). Other British goods will also cost U.S. consumers more, so that U.S. imports of British goods will decline.

From Britain's standpoint, the exchange rate (the pound price of dollars) has fallen (from £$\frac{1}{2}$ to £$\frac{1}{3}$ for $1). The international value of the pound has appreciated. The British previously got only $2 for £1; now they get $3 for £1. U.S. goods are therefore cheaper to the British, and U.S. exports to Britain will rise.

The two adjustments—a decrease in U.S. imports from Britain and an increase in U.S. exports to Britain—are just what are needed in terms of Figure 39.2 to decrease the quantity of pounds demanded from *b* to *c*, increase the quantity of pounds supplied from *a* to *c*, and thus correct the U.S. balance-of-payments deficit. These changes end when, at point *c*, the quantities of British pounds demanded and supplied are equal.

Disadvantages of Flexible Exchange Rates

Even though flexible exchange rates automatically work to eliminate payment imbalances, they may cause several significant problems. These are all related to the fact that flexible exchange rates are often volatile and can change by a large amount in just a few weeks or months. In addition, they often take substantial swings that can last several years or more. This can be seen in Figure 39.3, which plots the dollar-pound exchange rate from 1970 through 2012. (You can track other exchange rates, for example, the dollar-euro or dollar-yen rate, by going to the Federal Reserve Web site, **www.federalreserve.gov**, selecting Economic Research & Data, Statistical Releases and Historical Data, and finally, Exchange Rates and International Data.)

Uncertainty and Diminished Trade The risks and uncertainties associated with flexible exchange rates may discourage the flow of trade. Suppose a U.S. automobile dealer contracts to purchase 10 British cars for £150,000. At the current exchange rate of, say, $2 for £1, the U.S. importer expects to pay $300,000 for these automobiles. But if during the 3-month delivery period the rate of exchange shifts to $3 for £1, the £150,000 payment contracted by the U.S. importer will be $450,000.

That increase in the dollar price of pounds may thus turn the U.S. importer's anticipated profit into a substantial loss. Aware of the possibility of an adverse change in the exchange rate, the U.S. importer may not be willing to assume the risks involved. The U.S. firm may confine its operations to domestic automobiles, so international trade in this product will not occur.

The same thing can happen with investments. Assume that when the exchange rate is $3 to £1, a U.S. firm invests $30,000 (or £10,000) in a British enterprise. It estimates a return of 10 percent; that is, it anticipates annual earnings of $3,000 or £1,000. Suppose these expectations prove correct in that the British firm earns £1,000 in the first year on the £10,000 investment. But suppose that during the year, the value of the dollar appreciates to $2 = £1.

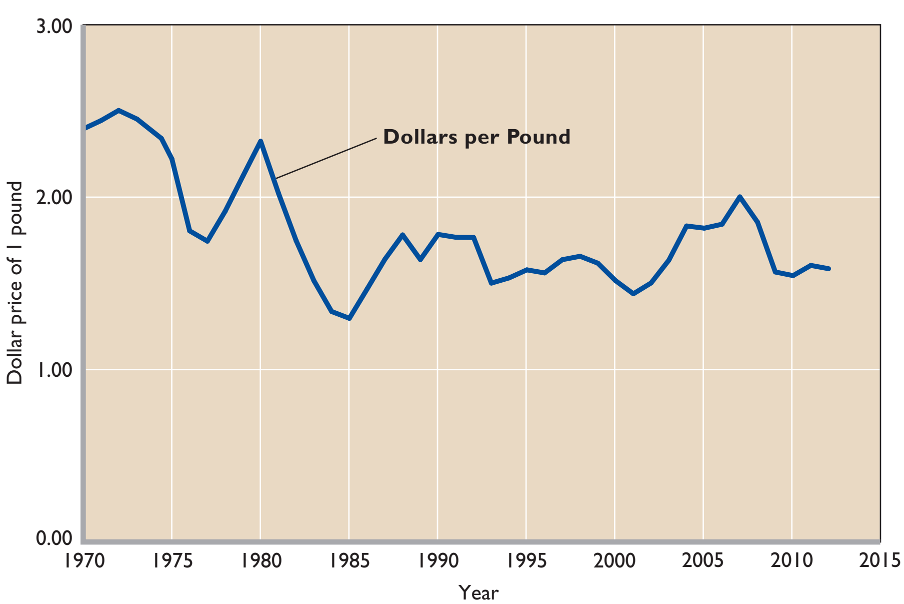


FIGURE 39.3 The dollar-pound exchange rate, 1970–2012. Before January 1971, the dollar-pound exchange rate was fixed at $2.40 = £1. Since that time, its value has been determined almost entirely by market forces, with only occasional government interventions. Under these mostly flexible conditions, the dollar-pound exchange rate has varied considerably. For instance, in 1981 it took $2.02 to buy a pound but by 1985 only $1.50 was needed to buy a pound. In contrast, between 2001 and 2007 the dollar price of a pound rose from $1.44 to $2.00, indicating that the dollar had depreciated relative to the pound. The dollar then sank back to $1.57 per British pound in 2009 and remained at about that level through 2012.

Source: Economic Report of the President, 2013, Table B-110. Earlier years from prior *Economic Reports*.

The absolute return is now only $2,000 (rather than $3,000), and the rate of return falls from the anticipated 10 percent to only $6\frac{2}{3}$ percent (= $2,000/$30,000). Investment is risky in any case. The added risk of changing exchange rates may persuade the U.S. investor not to venture overseas.[1]

Terms-of-Trade Changes A decline in the international value of its currency will worsen a nation's terms of trade. For example, an increase in the dollar price of a pound will mean that the United States must export more goods and services to finance a specific level of imports from Britain.

Instability Flexible exchange rates may destabilize the domestic economy because wide fluctuations stimulate and then depress industries producing exported goods. If the U.S. economy is operating at full employment and its currency depreciates, as in our illustration, the results will be inflationary, for two reasons. (1) Foreign demand for U.S. goods may rise, increasing total spending and pulling up U.S. prices. Also, the prices of all U.S. imports will increase. (2) Conversely, appreciation of the dollar will lower U.S. exports and increase imports, possibly causing unemployment.

Flexible or floating exchange rates also may complicate the use of domestic stabilization policies in seeking full employment and price stability. This is especially true for nations whose exports and imports are large relative to their total domestic output.

QUICK REVIEW 39.4

- Flexible exchange rates automatically adjust and eventually eliminate balance-of-payments deficits and surpluses.
- The volatility of flexible exchange rates may have several negative consequences, including discouraging international trade, worsening a nation's terms of trade, and destabilizing a nation's domestic economy by depressing export industries.

Fixed Exchange Rates

LO39.4 Describe the difference between flexible exchange rates and fixed exchange rates.

To circumvent the disadvantages of flexible exchange rates, at times nations have fixed or "pegged" their exchange rates. For our analysis of fixed exchange rates, we assume that the United States and Britain agree to maintain a $2 = £1 exchange rate.

The problem is that such a government agreement cannot keep from changing the demand for and the supply of pounds. With the rate fixed, a shift in demand or supply will threaten the fixed-exchange-rate system, and government must intervene to ensure that the exchange rate is maintained.

In Figure 39.2, suppose the U.S. demand for pounds increases from D_1 to D_2 and a U.S. payment deficit *ab* arises. Now, the new equilibrium exchange rate ($3 = £1) is above the fixed exchange rate ($2 = £1). How can the United States prevent the shortage of pounds from driving the exchange rate up to the new equilibrium level? How can it maintain the fixed exchange rate? The answer is by altering market demand or market supply or both so that they will intersect at the $2 = £1 rate. There are several ways to do this.

Use of Official Reserves

One way to maintain a fixed exchange rate is to engage in **currency interventions.** These are situations in which governments or their central banks manipulate an exchange rate through the use of official reserves. For instance, by selling some of its reserves of pounds, the U.S. government could increase the supply of pounds, shifting supply curve S_1 to the right so that it intersects D_2 at *b* in Figure 39.2, thereby maintaining the exchange rate at $2 = £1.

Notice that when the U.S. government sells some of its reserves of pounds, it is transferring assets to foreigners (since they gain ownership of the pounds). In terms of the balance-of-payments statement shown in Table 39.1, this transfer of assets enters positively on line 12, "Foreign purchases of assets in the United States." This positive entry is what offsets the balance-of-payments deficit caused by the fixed exchange rate and ensures that the U.S. balance of payments does in fact balance.

How do official reserves originate? Perhaps a balance-of-payments surplus occurred in the past. The U.S. government would have purchased that surplus. That is, at some earlier time, the U.S. government may have spent tax dollars to buy the surplus pounds that were threatening to reduce the exchange rate to below the $2 = £1 fixed rate. Those purchases would have bolstered the stock of U.S. official reserves of pounds.

Nations also have used gold as "international money" to obtain official reserves. In our example, the U.S. government could sell some of its gold to Britain to obtain pounds. It could then sell pounds for dollars. That would

[1] You will see in this chapter's Last Word, however, that a trader can circumvent part of the risk of unfavorable exchange-rate fluctuations by "hedging" in the "futures market" or "forward market" for foreign exchange.

shift the supply-of-pounds curve to the right, and the \$2 = £1 exchange rate could be maintained.

It is critical that the amount of reserves and gold be enough to accomplish the required increase in the supply of pounds. There is no problem if deficits and surpluses occur more or less randomly and are of similar size. Then, last year's balance-of-payments surplus with Britain will increase the U.S. reserve of pounds, and that reserve can be used to "finance" this year's deficit. But if the United States encounters persistent and sizable deficits for an extended period, it may exhaust its reserves, and thus be forced to abandon fixed exchange rates. Or, at the least, a nation whose reserves are inadequate must use less–appealing options to maintain exchange rates. Let's consider some of those options.

Trade Policies

To maintain fixed exchange rates, a nation can try to control the flow of trade and finance directly. The United States could try to maintain the \$2 = £1 exchange rate in the face of a shortage of pounds by discouraging imports (thereby reducing the demand for pounds) and encouraging exports (thus increasing the supply of pounds). Imports could be reduced by means of new tariffs or import quotas; special taxes could be levied on the interest and dividends U.S. financial investors receive from foreign investments. Also, the U.S. government could subsidize certain U.S. exports to increase the supply of pounds.

The fundamental problem is that these policies reduce the volume of world trade and change its makeup from what is economically desirable. When nations impose tariffs, quotas, and the like, they lose some of the economic benefits of a free flow of world trade. That loss should not be underestimated: Trade barriers by one nation lead to retaliatory responses from other nations, multiplying the loss.

Exchange Controls and Rationing

Another option is to adopt exchange controls and rationing. Under **exchange controls** the U.S. government could handle the problem of a pound shortage by requiring that all pounds obtained by U.S. exporters be sold to the federal government. Then the government would allocate or ration this short supply of pounds (represented by *xa* in Figure 39.2) among various U.S. importers, who demand the quantity *xb*. This policy would restrict the value of U.S. imports to the amount of foreign exchange earned by U.S. exports. Assuming balance in the capital and financial account, there would then be no balance-of-payments deficit. U.S. demand for British imports with the value *ab* would simply not be fulfilled.

There are major objections to exchange controls:

- ***Distorted trade*** Like *trade controls* (tariffs, quotas, and export subsidies), exchange controls would distort the pattern of international trade away from the pattern suggested by comparative advantage.
- ***Favoritism*** The process of rationing scarce foreign exchange might lead to government favoritism toward selected importers (big contributors to reelection campaigns, for example).
- ***Restricted choice*** Controls would limit freedom of consumer choice. The U.S. consumers who prefer Volkswagens might have to buy Chevrolets. The business opportunities for some U.S. importers might be impaired if the government were to limit imports.
- ***Black markets*** Enforcement problems are likely under exchange controls. U.S. importers might want foreign exchange badly enough to pay more than the \$2 = £1 official rate, setting the stage for black-market dealings between importers and illegal sellers of foreign exchange.

Domestic Macroeconomic Adjustments

A final way to maintain a fixed exchange rate would be to use domestic stabilization policies (monetary policy and fiscal policy) to eliminate the shortage of foreign currency. Tax hikes, reductions in government spending, and a high-interest-rate policy would reduce total spending in the U.S. economy and, consequently, domestic income. Because the volume of imports varies directly with domestic income, demand for British goods, and therefore for pounds, would be restrained.

If these "contractionary" policies served to reduce the domestic price level relative to Britain's, U.S. buyers of consumer and capital goods would divert their demands from British goods to U.S. goods, reducing the demand for pounds. Moreover, the high-interest-rate policy would lift U.S. interest rates relative to those in Britain.

Lower prices on U.S. goods and higher U.S. interest rates would increase British imports of U.S. goods and would increase British financial investment in the United States. Both developments would increase the supply of pounds. The combination of a decrease in the demand for and an increase in the supply of pounds would reduce or eliminate the original U.S. balance-of-payments deficit. In Figure 39.2 the new supply and demand curves would intersect at some new equilibrium point on line *ab*, where the exchange rate remains at \$2 = £1.

Maintaining fixed exchange rates by such means is hardly appealing. The "price" of exchange-rate stability for the United States would be a decline in output, employment, and price levels—in other words, a recession. Eliminating a balance-of-payments deficit and achieving domestic stability are both important national economic goals, but to sacrifice macroeconomic stability simply to balance international payments would be to let the tail wag the dog.

QUICK REVIEW 39.5

- To circumvent the disadvantages of flexible exchange rates, at times nations have fixed or "pegged" their exchange rates.
- Under a system of fixed exchange rates, nations set their exchange rates and then maintain them by buying or selling official reserves of currencies, establishing trade barriers, employing exchange controls, or incurring inflation or recession.

The Current Exchange Rate System: The Managed Float

LO39.5 Explain the current system of managed floating exchange rates.

Over the past 130 years, the world's nations have used three different exchange-rate systems. From 1879 to 1934, most nations used a gold standard, which implicitly created fixed exchange rates. From 1944 to 1971, most countries participated in the Bretton Woods system, which was a fixed-exchange-rate system indirectly tied to gold. And since 1971, most have used managed floating exchange rates, which mix mostly flexible exchange rates with occasional currency interventions. Naturally, our focus here is on the current exchange rate system. However, the history of the previous systems and why they broke down is highly fascinating. For that reason, we have included a discussion of these systems at the book's Web site (see Content Option for Instructors 2 [COI 2]).

The current international exchange-rate system (1971–present) is an "almost" flexible system called **managed floating exchange rates.** Exchange rates among major currencies are free to float to their equilibrium market levels, but nations occasionally use currency interventions in the foreign exchange market to stabilize or alter market exchange rates.

Normally, the major trading nations allow their exchange rates to float up or down to equilibrium levels based on supply and demand in the foreign exchange market. They recognize that changing economic conditions among nations require continuing changes in equilibrium exchange rates to avoid persistent payments deficits or surpluses. They rely on freely operating foreign exchange markets to accomplish the necessary adjustments. The result has been considerably more volatile exchange rates than those during the Bretton Woods era.

But nations also recognize that certain trends in the movement of equilibrium exchange rates may be at odds with national or international objectives. On occasion, nations therefore intervene in the foreign exchange market by buying or selling large amounts of specific currencies. This way, they can "manage" or stabilize exchange rates by influencing currency demand and supply.

The leaders of the *G8 nations* (Canada, France, Germany, Italy, Japan, Russia, United Kingdom, and United States) meet regularly to discuss economic issues and try to coordinate economic policies. At times they have collectively intervened to try to stabilize currencies. For example, in 2000 they sold dollars and bought euros in an effort to stabilize the falling value of the euro relative to the dollar. In the previous year the euro (€) had depreciated from €1 = \$1.17 to €1 = \$0.87.

The current exchange-rate system is thus an "almost" flexible exchange-rate system. The "almost" refers mainly to the occasional currency interventions by governments; it also refers to the fact that the actual system is more complicated than described. While the major currencies such as dollars, euros, pounds, and yen fluctuate in response to changing supply and demand, some developing nations peg their currencies to the dollar and allow their currencies to fluctuate with it against other currencies. Also, some nations peg the value of their currencies to a "basket" or group of other currencies.

How well has the managed float worked? It has both proponents and critics.

In Support of the Managed Float Proponents of the managed-float system argue that it has functioned far better than many experts anticipated. Skeptics had predicted that fluctuating exchange rates would reduce world trade and finance. But in real terms world trade under the managed float has grown tremendously over the past several decades. Moreover, as supporters are quick to point out, currency crises such as those in Mexico and southeast Asia in the last half of the 1990s were not the result of the floating-exchange-rate system itself. Rather, the abrupt currency devaluations and depreciations resulted from internal problems in those nations, in conjunction with the nations' tendency to peg their currencies to the dollar or to a basket of currencies.

In some cases, flexible exchange rates would have made these adjustments far more gradual.

Proponents also point out that the managed float has weathered severe economic turbulence that might have caused a fixed-rate system to break down. Such events as extraordinary oil price increases in 1973–1974 and again in 1981–1983, inflationary recessions in several nations in the mid-1970s, major national recessions in the early 1980s, and large U.S. budget deficits in the 1980s and the first half of the 1990s all caused substantial imbalances in international trade and finance, as did the large U.S. budget deficits and soaring world oil prices that occurred in the middle of the first decade of the 2000s. The U.S. financial crisis and the severe recession of 2007–2009 greatly disrupted world trade. Flexible rates enabled the system to adjust to all these events, whereas the same events would have put unbearable pressures on a fixed-rate system.

Concerns with the Managed Float There is still much sentiment in favor of greater exchange-rate stability. Those favoring more stable exchange rates see problems with the current system. They argue that the excessive volatility of exchange rates under the managed float threatens the prosperity of economies that rely heavily on exports. Several financial crises in individual nations (for example, Mexico, South Korea, Indonesia, Thailand, Russia, and Brazil) have resulted from abrupt changes in exchange rates. These crises have led to massive "bailouts" of those economies via loans from the International Monetary Fund (IMF). The IMF bailouts, in turn, may encourage nations to undertake risky and inappropriate economic policies since they know that, if need be, the IMF will come to the rescue. Moreover, some exchange-rate volatility has occurred even when underlying economic and financial conditions were relatively stable, suggesting that speculation plays too large a role in determining exchange rates.

Skeptics say the managed float is basically a "nonsystem" because the guidelines as to what each nation may or may not do with its exchange rates are not specific enough to keep the system working in the long run. Nations inevitably will be tempted to intervene in the foreign exchange market, not merely to smooth out short-term fluctuations in exchange rates but to prop up their currency if it is chronically weak or to manipulate the exchange rate to achieve domestic stabilization goals.

So what are we to conclude? Flexible exchange rates have not worked perfectly, but they have not failed miserably. Thus far they have survived, and no doubt have eased, several major shocks to the international trading system. Meanwhile, the "managed" part of the float has given nations some sense of control over their collective economic destinies. On balance, most economists favor continuation of the present system of "almost" flexible exchange rates.

QUICK REVIEW 39.6

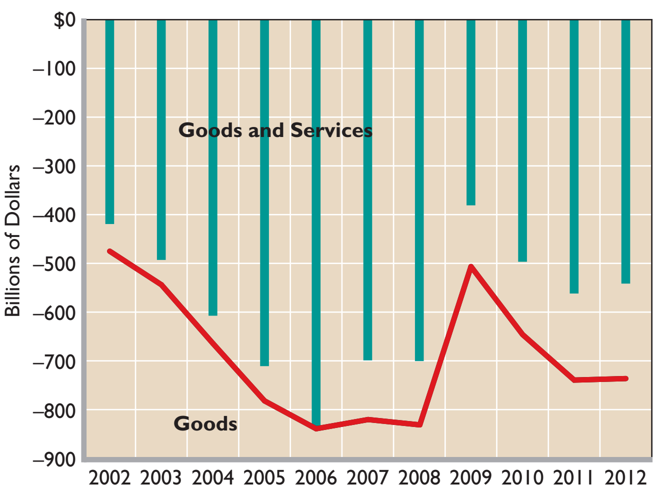
- The managed floating system of exchange rates (1971–present) relies on foreign exchange markets to establish equilibrium exchange rates.
- Under the system, nations can buy and sell official reserves of foreign currency to stabilize short-term changes in exchange rates or to correct exchange-rate imbalances that are negatively affecting the world economy.
- Proponents point out that international trade and investment have grown tremendously under the system. Critics say that it is a "nonsystem" and argue that the exchange rate volatility allowed under the managed float discourages international trade and investment. That is, trade and investment would be even larger if exchange rates were more stable.

Recent U.S. Trade Deficits

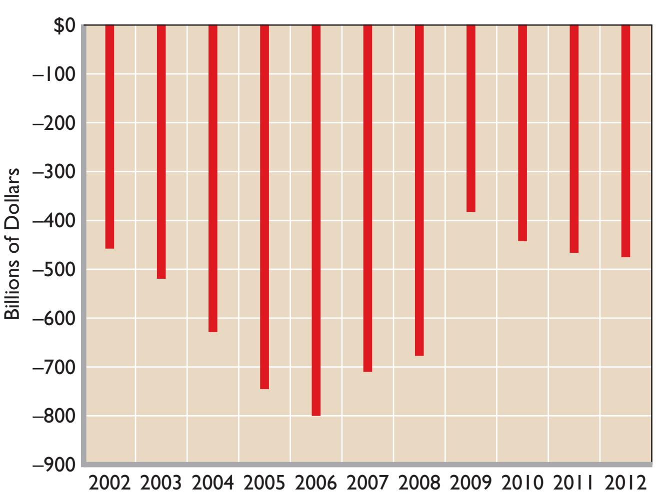
LO39.6 Identify the causes and consequences of recent U.S. trade deficits.

As shown in Figure 39.4a, the United States has experienced large and persistent trade deficits in recent years. These deficits rose rapidly between 2002 and 2006, with the trade deficit on goods and services peaking at $801 billion in 2006. The trade deficit on goods and services then declined precipitously to just $379 billion in 2009 as consumers and businesses greatly curtailed their purchases of imports during the recession of 2007–2009. As the economy recovered from the recession, the trade deficit on goods and services began rising again and reached $540 billion in 2012. The current account deficit (Figure 39.4b) reached a record high of $800 billion in 2006, and that amount was 6.0 percent of GDP. The current account deficit declined to $382 billion—2.6 percent of GDP—in the recession year 2009 before rebounding to $475 billion—or about 3.0 percent of GDP—in 2012. Economists expect the trade deficits to expand, absolutely and relatively, toward prerecession levels when the economy fully recovers and U.S. income and imports again rise.

FIGURE 39.4 U.S. trade deficits, 2002–2012. (a) The United States experienced large deficits in *goods* and in *goods and services* between 2002 and 2012. (b) The U.S. current account, generally reflecting the goods and services deficit, was also in substantial deficit. Although reduced significantly by the recession of 2007–2009, large current account deficits are expected to continue for many years to come.



(a)
Balance of trade



(b)
Balance on current accounts

Source: Bureau of Economic Analysis, **www.bea.gov**.

Causes of the Trade Deficits

The large U.S. trade deficits have several causes. First, the U.S. economy expanded more rapidly between 2002 and 2007 than the economies of several U.S. trading partners. The strong U.S. income growth that accompanied that economic growth enabled Americans to greatly increase their purchases of imported products. In contrast, Japan and some European nations suffered recession or experienced relatively slow income growth over that same period. So consumers in those countries increased their purchases of U.S. exports much less rapidly than Americans increased their purchases of foreign imports.

Another factor explaining the large trade deficits is the enormous U.S. trade imbalance with China. In 2012 the United States imported $298 billion more of goods and services than it exported to China. Even in the recession year 2009, the trade deficit with China was $220 billion. The 2012 deficit with China was 64 percent larger than the combined deficits with Mexico ($55 billion), Germany ($67 billion), and Japan ($60 billion). The United States is China's largest export market, and although China has greatly increased its imports from the United States, its standard of living has not yet risen sufficiently for its households to afford large quantities of U.S. products. Adding to the problem, China's government has fixed the exchange rate of it currency, the yuan, to a basket of currencies that includes the U.S. dollar. Therefore, China's large trade surpluses with the United States have not caused the yuan to appreciate much against the U.S. dollar. Greater appreciation of the yuan would have made Chinese goods more expensive in the United States and reduced U.S. imports from China. In China a stronger yuan would have reduced the dollar price of U.S. goods and increased Chinese purchases of U.S. exports. That combination—reduced U.S. imports from China and increased U.S. exports to China—would have reduced the large U.S. trade imbalance.

Another factor underlying the large U.S. trade deficits is a continuing trade deficit with oil-exporting nations. For example, in 2012 the United States had a $99 billion trade deficit with the OPEC countries.

A declining U.S. saving rate (= saving/total income) also contributed to the large U.S. trade deficits. Up until the recession of 2007–2009, the U.S. saving rate declined substantially, while its investment rate (= investment/total income) increased. The gap between U.S. investment and U.S. saving was filled by foreign purchases of U.S. real and financial assets, which created a large surplus on the U.S. capital and financial account. Because foreign savers were willing to finance a large part of U.S. investment, Americans were able to save less and consume more. Part of that added consumption spending was on imported goods.

Finally, many foreigners simply view U.S. assets favorably because of the relatively high risk-adjusted rates

LAST WORD

Speculation in Currency Markets

Are Speculators a Negative or a Positive Influence in Currency Markets and International Trade?

Most people buy foreign currency to facilitate the purchase of goods, services, or assets from another country. A U.S. importer buys Japanese yen to purchase Japanese autos. A Hong Kong financial investor purchases Australian dollars to invest in the Australian stock market. But there is another group of participants in the currency market—speculators—that buys and sells foreign currencies in the hope of reselling or rebuying them later at a profit.

Contributing to Exchange-Rate Fluctuations Speculators were much in the news in late 1997 and 1998 when they were widely accused of driving down the values of the South Korean won, Thai baht, Malaysian ringgit, and Indonesian rupiah. The value of these currencies fell by as much as 50 percent within one month, and speculators undoubtedly contributed to the swiftness of those declines. The expectation of currency depreciation (or appreciation) can be self-fulfilling. If speculators, for example, expect the Indonesian rupiah to be devalued or to depreciate, they quickly sell rupiah and buy currencies that they think will increase in relative value. The sharp increase in the supply of rupiah indeed reduces its value; this reduction then may trigger further selling of rupiah in expectation of further declines in its value.

But changed economic realities, not speculation, are normally the underlying causes of changes in currency values. That was largely the case with the southeast Asian countries in which actual and threatened bankruptcies in the financial and manufacturing sectors undermined confidence in the strength of the currencies. Anticipating the eventual declines in currency values, speculators simply hastened those declines. That is, the declines in value probably would have occurred with or without speculators.

Moreover, on a daily basis, speculation clearly has positive effects in foreign exchange markets.

of return they provide. The purchase of those assets provides foreign currency to Americans that enables them to finance their strong appetite for imported goods. The capital account surpluses therefore may partially cause the high U.S. trade deficits, not just result from those high deficits. The point is that the causes of the high U.S. trade deficits are numerous and not so easy to disentangle.

Implications of U.S. Trade Deficits

The prerecession U.S. trade deficits were the largest ever run by a major industrial nation. Whether the large trade deficits should be of significant concern to the United States and the rest of the world is debatable. Most economists see both benefits and costs to trade deficits.

Increased Current Consumption At the time a trade deficit or a current account deficit is occurring, American consumers benefit. A trade deficit means that the United States is receiving more goods and services as imports from abroad than it is sending out as exports. Taken alone, a trade deficit allows the United States to consume outside its production possibilities curve. It augments the domestic standard of living. But here is a catch: The gain in present consumption may come at the expense of reduced future consumption. When and if the current account deficit declines, Americans may have to consume less than before and perhaps even less than they produce.

Increased U.S. Indebtedness A trade deficit is considered unfavorable because it must be financed by

Smoothing Out Short-Term Fluctuations in Currency Prices When temporarily weak demand or strong supply reduces a currency's value, speculators quickly buy the currency, adding to its demand and strengthening its value. When temporarily strong demand or weak supply increases a currency's value, speculators sell the currency. That selling increases the supply of the currency and reduces its value. In this way speculators smooth out supply and demand, and thus exchange rates, over short time periods. This day-to-day exchange-rate stabilization aids international trade.

Absorbing Risk Speculators also absorb risk that others do not want to bear. Because of potential adverse changes in exchange rates, international transactions are riskier than domestic transactions. Suppose AnyTime, a hypothetical retailer, signs a contract with a Swiss manufacturer to buy 10,000 Swatch watches to be delivered in three months. The stipulated price is 75 Swiss francs per watch, which in dollars is \$50 per watch at the present exchange rate of, say, \$1 = 1.5 francs. AnyTime's total bill for the 10,000 watches will be \$500,000 (= 750,000 francs).

But if the Swiss franc were to appreciate, say, to \$1 = 1 franc, the dollar price per watch would rise from \$50 to \$75 and AnyTime would owe \$750,000 for the watches (= 750,000 francs). AnyTime may reduce the risk of such an unfavorable exchange-rate fluctuation by hedging in the futures market. Hedging is an action by a buyer or a seller to protect against a change in future prices. The futures market is a market in which currencies are bought and sold at prices fixed now, for delivery at a specified date in the future.

AnyTime can purchase the needed 750,000 francs at the current \$1 = 1.5 francs exchange rate, but with delivery in three months when the Swiss watches are delivered. And here is where speculators come in. For a price determined in the futures market, they agree to deliver the 750,000 francs to AnyTime in three months at the \$1 = 1.5 francs exchange rate, regardless of the exchange rate then. The speculators need not own francs when the agreement is made. If the Swiss franc depreciates to, say, \$1 = 2 francs in this period, the speculators profit. They can buy the 750,000 francs stipulated in the contract for \$375,000, pocketing the difference between that amount and the \$500,000 AnyTime has agreed to pay for the 750,000 francs. If the Swiss franc appreciates, the speculators, but not AnyTime, suffer a loss.

The amount AnyTime must pay for this "exchange-rate insurance" will depend on how the market views the likelihood of the franc depreciating, appreciating, or staying constant over the three-month period. As in all competitive markets, supply and demand determine the price of the futures contract.

The futures market thus eliminates much of the exchange-rate risk associated with buying foreign goods for future delivery. Without it, AnyTime might have decided against importing Swiss watches. But the futures market and currency speculators greatly increase the likelihood that the transaction will occur. Operating through the futures market, speculation promotes international trade.

In short, although speculators in currency markets occasionally contribute to swings in exchange rates, on a day-to-day basis they play a positive role in currency markets.

borrowing from the rest of the world, selling off assets, or dipping into official reserves. Recall that current account deficits are financed by surpluses in the capital and financial accounts. Such surpluses require net inpayments of dollars to buy U.S. assets, including debt issued by Americans. Therefore, when U.S. exports are insufficient to finance U.S. imports, the United States increases both its debt to people abroad and the value of foreign claims against assets in the United States. Financing of the U.S. trade deficit has resulted in a larger foreign accumulation of claims against U.S. financial and real assets than the U.S. claim against foreign assets. In 2008, foreigners owned about \$3.5 trillion more of U.S. assets (corporations, land, stocks, bonds, loan notes) than U.S. citizens and institutions owned in foreign assets.

If the United States wants to regain ownership of these domestic assets, at some future time it will have to export more than it imports. At that time, domestic consumption will be lower because the United States will need to send more of its output abroad than it receives as imports. Therefore, the current consumption gains delivered by U.S. current account deficits may mean permanent debt, permanent foreign ownership, or large sacrifices of future consumption.

We say "may mean" above because the foreign lending to U.S. firms and foreign investment in the United States increases the U.S. capital stock. U.S. production capacity therefore might increase more rapidly than other-wise because of a large surplus on the capital and financial account. Faster increases in production capacity and real GDP enhance the economy's ability to service foreign debt and buy back real capital, if that is desired.

Trade deficits therefore are a mixed blessing. The long-term impacts of the record-high U.S. trade deficits are largely unknown. That "unknown" worries some economists, who are concerned that foreigners will lose financial confidence in the United States. If that happens, they would restrict their lending to American households and businesses and also reduce their purchases of U.S. assets. Both actions would decrease the demand for U.S. dollars in the foreign exchange market and cause the U.S. dollar to depreciate. A sudden, large depreciation of the U.S. dollar might disrupt world trade and negatively affect economic growth worldwide. Other economists, however, downplay this scenario. Because any decline in the U.S. capital and financial account surplus is automatically met with a decline in the current account deficit, U.S. net exports would rise and the overall impact on the American economy would be slight.

QUICK REVIEW 39.7

- The United States has had large trade deficits in recent decades.
- Causes include (a) more rapid income growth in the United States than in Japan and some European nations, resulting in expanding U.S. imports relative to exports; (b) the emergence of a large trade deficit with China; (c) continuing large trade deficits with oil-exporting nations; and (d) a large surplus in the capital and financial account, which enabled Americans to reduce their saving and buy more imports.
- The severe recession of 2007–2009 in the United States substantially lowered the U.S. trade deficit by reducing American spending on imports.
- U.S. trade deficits have produced current increases in the living standards of U.S. consumers but the accompanying surpluses on the capital and financial account have increased U.S. debt to the rest of the world and increased foreign ownership of assets in the United States.

SUMMARY

LO39.1 Explain how currencies of different nations are exchanged when international transactions take place.

International financial transactions involve trade either in currently produced goods and services or in preexisting assets. Exports of goods, services, and assets create inflows of money, while imports cause outflows of money. If buyers and sellers use different currencies, then foreign exchange transactions take place so that the exporter can be paid in his or her own currency.

LO39.2 Analyze the balance sheet the United States uses to account for the international payments it makes and receives.

The balance of payments records all international trade and financial transactions taking place between a given nation and the rest of the world. The balance on goods and services (the trade balance) compares exports and imports of both goods and services. The current account balance includes not only goods and services transactions but also net investment income and net transfers.

The capital and financial account includes (*a*) the net amount of the nation's debt forgiveness and (*b*) the nation's sale of real and financial assets to people living abroad less its purchases of real and financial assets from foreigners.

The current account and the capital and financial account always sum to zero. A deficit in the current account is always offset by a surplus in the capital and financial account. Conversely, a surplus in the current account is always offset by a deficit in the capital and financial account.

Official reserves are owned by national governments and their central banks and consist of stocks of foreign currencies, certain reserves held with the International Monetary Fund, and stocks of gold.

A balance-of-payments deficit is said to occur when a nation draws down its stock of official reserves to purchase dollars from abroad to balance the capital and financial account with the current account. A balance-of-payments surplus occurs when a nation adds to its stock of official reserves by selling dollars to foreigners to obtain foreign currencies to balance the two accounts. The desirability of a balance-of-payments deficit or surplus depends on its size and its persistence.

LO39.3 Discuss how exchange rates are determined in currency markets that have flexible exchange rates.

Flexible or floating exchange rates between international currencies are determined by the demand for and supply of those currencies. Under flexible rates, a currency will depreciate or appreciate as a result of changes in tastes, relative income changes, relative changes in inflation rates, relative changes in real interest rates, and speculation.

LO39.4 Describe the difference between flexible exchange rates and fixed exchange rates.

The maintenance of fixed exchange rates requires adequate official reserves to accommodate periodic payments deficits. If reserves are inadequate, nations must invoke protectionist trade policies, engage in exchange controls, or endure undesirable domestic macroeconomic adjustments.

LO39.5 Explain the current system of managed floating exchange rates.

Since 1971 the world's major nations have used a system of managed floating exchange rates. Market forces generally set rates, although governments intervene with varying frequency to alter their exchange rates.

LO39.6 Identify the causes and consequences of recent U.S. trade deficits.

Between 1997 and 2007, the United States had large and rising trade deficits, which are projected to last well into the future. Causes of the trade deficits include (*a*) more rapid income growth in the United States than in Japan and some European nations, resulting in expanding U.S. imports relative to exports, (*b*) the emergence of a large trade deficit with China, (*c*) continuing large trade deficits with oil-exporting nations, and (*d*) a large surplus in the capital and financial account, which enabled Americans to reduce their saving and buy more imports. The severe recession of 2007–2009 in the United States substantially lowered the U.S. trade deficit by reducing American spending on imports.

U.S. trade deficits have produced current increases in the living standards of U.S. consumers. The accompanying surpluses on the capital and financial account have increased U.S. debt to the rest of the world and increased foreign ownership of assets in the United States. This greater foreign investment in the United States, however, has undoubtedly increased U.S. production possibilities.

TERMS AND CONCEPTS

balance of payments
current account
balance on goods and services
trade deficit
trade surplus
balance on current account
capital and financial account
balance on capital and financial account
official reserves
balance-of-payments deficits and surpluses
flexible- or floating-exchange-rate system
fixed-exchange-rate system
purchasing-power-parity theory
currency interventions
exchange controls
managed floating exchange rates

The following and additional problems can be found in connect ECONOMICS

DISCUSSION QUESTIONS

1. Do all international financial transactions necessarily involve exchanging one nation's distinct currency for another? Explain. Could a nation that neither imports goods and services nor exports goods and services still engage in international financial transactions? **LO39.1**
2. Explain: "U.S. exports earn supplies of foreign currencies that Americans can use to finance imports." Indicate whether each of the following creates a demand for or a supply of European euros in foreign exchange markets: **LO39.1**
   a. A U.S. airline firm purchases several Airbus planes assembled in France.
   b. A German automobile firm decides to build an assembly plant in South Carolina.
   c. A U.S. college student decides to spend a year studying at the Sorbonne in Paris.
   d. An Italian manufacturer ships machinery from one Italian port to another on a Liberian freighter.
   e. The U.S. economy grows faster than the French economy.
   f. A U.S. government bond held by a Spanish citizen matures, and the loan amount is paid back to that person.
   g. It is widely expected that the euro will depreciate in the near future.
3. What do the plus signs and negative signs signify in the U.S. balance-of-payments statement? Which of the following items appear in the current account and which appear in the capital and financial account? U.S. purchases of assets abroad; U.S. services imports; foreign purchases of assets in the United States; U.S. goods exports, U.S. net investment income. Why must the current account and the capital and financial account sum to zero? **LO39.2**
4. What are official reserves? How do net sales of official reserves to foreigners and net purchases of official reserves from foreigners relate to U.S. balance-of-payments deficits and surpluses? Explain why these deficits and surpluses are not actual deficits and surpluses in the *overall* balance-of-payments statement. **LO39.2**
5. Generally speaking, how is the dollar price of euros determined? Cite a factor that might increase the dollar price of euros. Cite a different factor that might decrease the dollar price of euros. Explain: "A rise in the dollar price of euros necessarily means a fall in the euro price of dollars." Illustrate and elaborate: "The dollar-euro exchange rate provides a direct link between the prices of goods and services produced in the eurozone and in the United States." Explain the

purchasing-power-parity theory of exchange rates, using the euro-dollar exchange rate as an illustration. **LO39.3**

6. Suppose that a Swiss watchmaker imports watch components from Sweden and exports watches to the United States. Also suppose the dollar depreciates, and the Swedish krona appreciates, relative to the Swiss franc. Speculate as to how each would hurt the Swiss watchmaker. **LO39.3**
7. Explain why the U.S. demand for Mexican pesos is downsloping and the supply of pesos to Americans is upsloping. Assuming a system of flexible exchange rates between Mexico and the United States, indicate whether each of the following would cause the Mexican peso to appreciate or depreciate, other things equal: **LO39.3**
   - **a.** The United States unilaterally reduces tariffs on Mexican products.
   - **b.** Mexico encounters severe inflation.
   - **c.** Deteriorating political relations reduce American tourism in Mexico.
   - **d.** The U.S. economy moves into a severe recession.
   - **e.** The United States engages in a high-interest-rate monetary policy.
   - **f.** Mexican products become more fashionable to U.S. consumers.
   - **g.** The Mexican government encourages U.S. firms to invest in Mexican oil fields.
   - **h.** The rate of productivity growth in the United States diminishes sharply.
8. Explain why you agree or disagree with the following statements. Assume other things equal. **LO39.3**
   - **a.** A country that grows faster than its major trading partners can expect the international value of its currency to depreciate.
   - **b.** A nation whose interest rate is rising more rapidly than interest rates in other nations can expect the international value of its currency to appreciate.
   - **c.** A country's currency will appreciate if its inflation rate is less than that of the rest of the world.
9. "Exports pay for imports. Yet in 2012 the nations of the world exported about $540 billion more of goods and services to the United States than they imported from the United States." Resolve the apparent inconsistency of these two statements. **LO39.2**
10. What have been the major causes of the large U.S. trade deficits in recent years? What are the major benefits and costs associated with trade deficits? Explain: "A trade deficit means that a nation is receiving more goods and services from abroad than it is sending abroad." How can that be considered to be "unfavorable"? **LO39.6**
11. **LAST WORD** Suppose Super D'Hiver—a hypothetical French snowboard retailer—wants to order 5,000 snowboards made in the United States. The price per board is $200, the present exchange rate is 1 euro = $1, and payment is due in dollars when the boards are delivered in 3 months. Use a numerical example to explain why exchange-rate risk might make the French retailer hesitant to place the order. How might speculators absorb some of Super D'Hiver's risk?

REVIEW QUESTIONS

1. An American company wants to buy a television from a Chinese company. The Chinese company sells its TVs for 1,200 yuan each. The current exchange rate between the U.S. dollar and the Chinese yuan is $1 = 6 yuan. How many dollars will the American company have to convert into yuan to pay for the television? **LO39.1**
   - **a.** $7,200.
   - **b.** $1,200.
   - **c.** $200.
   - **d.** $100.
2. Suppose that a country has a trade surplus of $50 billion, a balance on the capital account of $10 billion, and a balance on the current account of −$200 billion. The balance on the capital and financial account will be: **LO39.2**
   - **a.** $10 billion.
   - **b.** $50 billion.
   - **c.** $200 billion.
   - **d.** −$200 billion.
3. The exchange rate between the U.S. dollar and the British pound starts at $1 = £0.5. It then changes to $1 = £0.75. Given this change, we would say that the U.S. dollar has ____________ while the British pound has ____________. **LO39.3**
   - **a.** Depreciated; appreciated.
   - **b.** Depreciated; depreciated.
   - **c.** Appreciated; depreciated.
   - **d.** Appreciated; appreciated.
4. A meal at a McDonald's restaurant in New York costs $8. The identical meal at a McDonald's restaurant in London costs £4. According to the purchasing-power-parity theory of exchange rates, the exchange rate between U.S. dollars and British pounds should tend to move toward: **LO39.3**
   - **a.** $2 = £1.
   - **b.** $1 = £2.
   - **c.** $4 = £1.
   - **d.** $1 = £4.
5. Suppose that a country has a flexible exchange rate. Also suppose that at the current exchange rate, the country is experiencing a balance-of-payments deficit. Then would it be true or false that a sufficiently large depreciation of the local currency could eliminate the balance-of-payments deficit. **LO39.3**
6. Diagram a market in which the equilibrium dollar price of 1 unit of fictitious currency zee (Z) is $5 (the exchange rate is $5 = Z1). Then show on your diagram a decline in the demand for zee. **LO39.4**

a. Referring to your diagram, discuss the adjustment options the United States would have in maintaining the exchange rate at $5 = Z1 under a fixed-exchange-rate system.

b. How would the U.S. balance-of-payments surplus that is caused by the decline in demand be resolved under a system of flexible exchange rates?

7. Suppose that the government of China is currently fixing the exchange rate between the U.S. dollar and the Chinese yuan at a rate of $1 = 6 yuan. Also suppose that at this exchange rate, the people who want to convert dollars to yuan are asking to convert $10 billion per day of dollars into yuan, while the people who are wanting to convert yuan into dollars are asking to convert 36 billion yuan into dollars. What will happen to the size of China's official reserves of dollars? **LO39.4**

a. Increase.
b. Decrease.
c. Stay the same.

8. Suppose that a country follows a managed-float policy but that its exchange rate is currently floating freely. In addition, suppose that it has a massive current account deficit. Does it also necessarily have a balance-of-payments deficit? If it decides to engage in a currency intervention to reduce the size of its current account deficit, will it buy or sell its own currency? As it does so, will its official reserves of foreign currencies get larger or smaller? Would that outcome indicate a balance-of-payments deficit or a balance-of-payments surplus? **LO39.5**

9. If the economy booms in the United States while going into recession in other countries, the U.S. trade deficit will tend to ____________. **LO39.6**

a. Increase.
b. Decrease.
c. Remain the same.

10. Other things equal, if the United States continually runs trade deficits, foreigners will own ____________ U.S. assets. **LO39.6**

a. More and more.
b. Less and less.
c. The same amount of.

PROBLEMS

1. Alpha's balance-of-payments data for 2012 are shown below. All figures are in billions of dollars. What are the (*a*) balance on goods, (*b*) balance on goods and services, (*c*) balance on current account, and (*d*) balance on capital and financial account? Suppose Alpha sold $10 billion of official reserves abroad to balance the capital and financial account with the current account. Does Alpha have a balance-of-payments deficit or does it have a surplus? **LO39.2**

Goods exports	$+40
Goods imports	−30
Service exports	+15
Service imports	−10
Net investment income	−5
Net transfers	+10
Balance on capital account	0
Foreign purchases of Alpha assets	+20
Alpha purchases of assets abroad	−40

2. China had a $214 billion overall current account surplus in 2012. Assuming that China's net debt forgiveness was zero in 2012 (its capital account balance was zero), by how much did Chinese purchases of financial and real assets abroad exceed foreign purchases of Chinese financial and real assets? **LO39.2**

3. Refer to the following table, in which Q_d is the quantity of loonies demanded, P is the dollar price of loonies, Q_s is the quantity of loonies supplied in year 1, and Q_s' is the quantity of loonies supplied in year 2. All quantities are in billions and the dollar-loonie exchange rate is fully flexible. **LO39.3**

Q_d	P	Q_s	Q_s'
10	125	30	20
15	120	25	15
20	115	20	10
25	110	15	5

a. What is the equilibrium dollar price of loonies in year 1?
b. What is the equilibrium dollar price of loonies in year 2?
c. Did the loonie appreciate or did it depreciate relative to the dollar between years 1 and 2?
d. Did the dollar appreciate or did it depreciate relative to the loonie between years 1 and 2?
e. Which one of the following could have caused the change in relative values of the dollar (used in the United States) and the loonie (used in Canada) between years 1 and 2: (1) More rapid inflation in the United States than in Canada, (2) an increase in the real interest rate in the United States but not in Canada, or (3) faster income growth in the United States than in Canada.

4. Suppose that the current Canadian dollar (CAD) to U.S. dollar exchange rate is $0.85 CAD = $1 US and that the U.S. dollar price of an Apple iPhone is $300. What is the Canadian dollar price of an iPhone? Next, suppose that the CAD to U.S. dollar exchange rate moves to $0.96 CAD = $1 US. What is the new Canadian dollar price of an iPhone? Other things equal, would you expect Canada to import more or fewer iPhones at the new exchange rate? **LO39.3**

5. Return to problem 3 and assume the exchange rate is fixed against the dollar at the equilibrium exchange rate that occurs in year 1. Also suppose that Canadia and the United States are the only two countries in the world. In year 2, what quantity of loonies would the government of Canadia have to buy or sell to balance its capital and financial account with its current account? In what specific account would this purchase or sale show up in Canadia's balance-of-payments statement: Foreign purchases of assets in Canadia or Canadia's purchases of assets abroad? Would this transaction increase Canadia's stock of official reserves or decrease its stock? **LO39.6**

FURTHER TEST YOUR KNOWLEDGE AT www.mcconnell20e.com

Practice quizzes, student PowerPoints, worked problems, Web-based questions, and additional materials are available at the text's Online Learning Center (OLC), **www.mcconnell20e.com**, or scan here. Need a barcode reader? Try ScanLife, available in your app store.

WEB CHAPTER 39

The Economics of Developing Countries

www.mcconnell20.com

Learning Objectives

LO39W.1 Describe how the World Bank distinguishes between industrially advanced countries (high-income nations) and developing countries (middle-income and low-income nations).

LO39W.2 List some of the obstacles to economic development.

LO39W.3 Explain the vicious circle of poverty that afflicts low-income nations.

LO39W.4 Discuss the role of government in promoting economic development within low-income nations.

WEB CHAPTER

LO39W.5 Describe how industrial nations attempt to aid low-income countries.

It is difficult for those of us in the United States, where per capita GDP in 2012 was about $42,683, to grasp the fact that about 2.5 billion people, or nearly half the world's population, live on $2 or less a day. And about 1.3 billion live on less than $1.25 a day. Hunger, squalor, and disease are the norm in many nations of the world.

In this bonus Web chapter (at our Web site, **www.mcconnell20e.com**), we identify the developing countries, discuss their characteristics, and explore the obstacles that have impeded their growth. We also examine the appropriate roles of the private sector and government in economic development. Finally, we look at policies that might help developing countries increase their growth rates.

GLOSSARY

Note: Terms set in *italic* type are defined separately in this glossary.

ability-to-pay principle The idea that those who have greater *income* (or *wealth*) should pay a greater proportion of it as taxes than those who have less income (or wealth).

absolute advantage A situation in which a person or country can produce more of a particular product from a specific quantity of resources than some other person or country.

accounting profit The *total revenue* of a *firm* less its *explicit costs;* the profit (or net income) that appears on accounting statements and that is reported to the government for tax purposes.

acreage allotments A pre-1996 government program that limited the total number of acres to be used in producing (reduced amounts of) various food and fiber products and allocated these acres among individual farmers. These farmers had to limit their plantings to the allotted number of acres to obtain *price supports* for their crops.

actively managed funds *Mutual funds* that have portfolio managers who constantly buy and sell *assets* in an attempt to generate high returns.

actual investment The amount that *firms* invest; equal to *planned investment* plus *unplanned investment* (which may be negative).

actual reserves The funds that a bank has on deposit at the *Federal Reserve Bank* of its district (plus its *vault cash*).

adverse selection problem A problem arising when information known to one party to a contract or agreement is not known to the other party, causing the latter to incur major costs. Example: Individuals who have the poorest health are most likely to buy health insurance.

advertising A seller's activities in communicating its message about its product to potential buyers.

AFL-CIO An acronym for the American Federation of Labor–Congress of Industrial Organizations; the largest federation of *labor unions* in the United States.

agency shop A place of employment where the employer may hire either *labor union* members or nonmembers but where those employees who do not join the union must either pay union dues or donate an equivalent amount of money to a charity.

aggregate A collection of specific economic units treated as if they were one unit. Examples: the *prices* of all individual *goods* and *services* are combined into the *price level*, and all units of output are aggregated into *gross domestic product*.

aggregate demand A schedule or curve that shows the total quantity of *goods* and *services* that would be demanded (purchased) at various *price levels*.

aggregate demand–aggregate supply (AD-AS) model The macroeconomic model that uses *aggregate demand* and *aggregate supply* to determine and explain the *price level* and the real *domestic output* (*real gross domestic product*).

aggregate expenditures The total amount spent for *final goods* and final *services* in an economy.

aggregate expenditures–domestic output approach Determination of the equilibrium *gross domestic product* by finding the real GDP at which *aggregate expenditures* equal *domestic output*.

aggregate expenditures model The *macroeconomics* model developed by John Maynard Keynes that assumes completely *inflexible prices*, thereby forcing the economy to adjust toward an *equilibrium real domestic output* by a process in which *firms*, reacting to unexpected changes in *inventory* levels, adjust the volume of output until the *aggregate expenditures* made on *final goods* and final *services* just equal the amount of output being produced in the economy.

aggregate expenditures schedule A table of numbers showing the total amount spent on *final goods* and final *services* at different levels of *real gross domestic product* (*real GDP*).

aggregate supply A schedule or curve showing the total quantity of *goods* and *services* that would be supplied (produced) at various *price levels*.

aggregate supply shocks Sudden, large changes in resource costs that shift an economy's aggregate supply curve.

agribusiness The portion of the agricultural and food product industries that is dominated by large corporations.

Alcoa case A 1945 case in which the courts ruled that the possession of monopoly power, no matter how reasonably that power had been used, was a violation of the antitrust laws; temporarily overturned the *rule of reason* applied in the *U.S. Steel case*.

allocative efficiency The apportionment of resources among *firms* and industries to obtain the production of the products most wanted by society (consumers); the output of each product at which its *marginal cost* and *price* or *marginal benefit* are equal, and at which the sum of *consumer surplus* and *producer surplus* is maximized.

anchoring The tendency people have to unconsciously base, or "anchor," the valuation of an item they are currently thinking about on recently considered but logically irrelevant information.

anticipated inflation Increases in the *price level* (*inflation*) that occur at the expected rate.

antitrust laws Legislation (including the *Sherman Act* and *Clayton Act*) that prohibits anticompetitive business activities such as *price fixing*, bid rigging, monopolization, and *tying contracts*.

antitrust policy The use of the *antitrust laws* to promote *competition* and *economic efficiency*.

appreciation (of the dollar) An increase in the value of the dollar relative to the currency of another nation, so a dollar buys a larger amount of the foreign currency and thus of foreign goods.

arbitrage The activity of selling one *asset* and buying an identical or nearly identical asset to benefit from temporary differences in *prices* or *rates of return;* the practice that equalizes prices or returns on similar financial instruments and thus eliminates further opportunities for riskless financial gains.

asset Anything of monetary value owned by a *firm* or individual.

asset demand for money The amount of *money* people want to hold as a *store of value;* this amount varies inversely with the *interest rate*.

asymmetric information A situation where one party to a market transaction has much more information about a product or service than the other. The result may be an under- or overallocation of resources.

average expected rate of return The *probability-weighted average* of an investment's possible future returns.

average fixed cost (AFC) A *firm*'s total *fixed cost* divided by output (the quantity of product produced).

average product (AP) The total output produced per unit of a *resource* employed (*total product* divided by the quantity of that employed resource).

average propensity to consume (APC) Fraction (or percentage) of *disposable income* that *households* spend on *consumer goods;* consumption divided by *disposable income*.

average propensity to save (APS) Fraction (or percentage) of *disposable income* that *households* save; *saving* divided by *disposable income*.

average revenue Total revenue from the sale of a product divided by the quantity of the product sold (demanded); equal to the *price* at which the product is sold when all units of the product are sold at the same price.

average tax rate Total tax paid divided by total *taxable income* or some other base (such as total income) against which to compare the amount of tax paid. Expressed as a percentage.

average total cost (ATC) A firm's *total cost* divided by output (the quantity of product produced); equal to *average fixed cost* plus *average variable cost*.

average variable cost (AVC) A firm's total *variable cost* divided by output (the quantity of product produced).

backflows The return of workers to the countries from which they originally emigrated.

balance of payments A summary of all the financial transactions that take place between the individuals, *firms*, and governmental units of one nation and those of all other nations during a year.

balance-of-payments deficit The net amount of *official reserves* (mainly foreign currencies) that a nation's treasury or central bank must sell to achieve balance between that nation's *capital and financial account* and its *current account* (in its *balance of payments*).

balance-of-payments surplus The net amount of *official reserves* (mainly foreign currencies) that a nation's treasury or central bank must buy to achieve balance between that nation's *capital and financial account* and its *current account* (in its *balance of payments*).

balance on capital and financial account The sum of the *capital account balance* and the *financial account balance*.

balance on current account The exports of *goods* and *services* of a nation less its imports of goods and services plus its *net investment income* and *net transfers* in a year.

balance on goods and services The exports of *goods* and *services* of a nation less its imports of goods and services in a year.

balance sheet A statement of the *assets*, *liabilities*, and *net worth* of a *firm* or individual at some given time.

bank deposits The deposits that individuals or *firms* have at banks (or thrifts) or that banks have at the *Federal Reserve Banks*.

bankers' bank A bank that accepts the deposits of and makes loans to *depository institutions;* in the United States, a *Federal Reserve Bank*.

bank reserves The deposits of commercial banks and thrifts at *Federal Reserve Banks* plus bank and thrift *vault cash*.

bankrupt A legal situation in which an individual or *firm* finds that it cannot make timely interest payments on money it has borrowed. In such cases, a bankruptcy judge can order the individual or firm to liquidate (turn into cash) its assets in order to pay lenders at least some portion of the amount they are owed.

barrier to entry Anything that artificially prevents the entry of *firms* into an *industry*.

barter The direct exchange of one *good* or *service* for another good or service.

base year The year with which other years are compared when an index is constructed; for example, the base year for a *price index*.

beaten paths Migration routes taken previously by family, relatives, friends, and other migrants.

behavioral economics The branch of economic theory that combines insights from economics, psychology, and biology to make more accurate predictions about human behavior than conventional *neoclassical economics*, which is hampered by its core assumptions that people are fundamentally *rational* and almost entirely self-interested. Behavioral economics can explain *framing effects*, *anchoring*, *mental accounting*, the *endowment effect*, *status quo bias*, *time inconsistency*, and *loss aversion*.

benefits-received principle The idea that those who receive the benefits of *goods* and *services* provided by government should pay the taxes required to finance them.

beta A relative measure of *nondiversifiable risk* that measures how the nondiversifiable risk of a given *asset* or *portfolio* compares with that of the *market portfolio* (the portfolio that contains every asset available in the financial markets).

bilateral monopoly A market in which there is a single seller (*monopoly*) and a single buyer (*monopsony*).

Board of Governors The seven-member group that supervises and controls the money and banking system of the United States; the Board of Governors of the *Federal Reserve System;* the Federal Reserve Board.

bond A financial device through which a borrower (a *firm* or government) is obligated to pay the principal and interest on a loan at specific dates in the future.

brain drains The exit or *emigration* of highly educated, highly skilled workers from a country.

break-even income The level of *disposable income* at which *households* plan to consume (spend) all their income and to save none of it.

break-even output Any output at which a (competitive) *firm*'s *total cost* and *total revenue* are equal; an output at which a firm has neither an *economic profit* nor an economic loss, at which it earns only a *normal profit*.

break-even point An output at which a *firm* makes a *normal profit* (*total revenue* = *total cost*) but not an *economic profit*.

British thermal unit (BTU) The amount of energy required to raise the temperature of 1 pound of water by 1 degree Fahrenheit.

budget constraint The limit that the size of a consumer's income (and the *prices* that must be paid for *goods* and *services*) imposes on the ability of that consumer to obtain goods and services.

budget deficit The amount by which expenditures exceed revenues in any year.

budget line A line that shows the different combinations of two products a consumer can purchase with a specific money income, given the products' *prices*.

budget surplus The amount by which the revenues of the federal government exceed its expenditures in any year.

built-in stabilizer A mechanism that increases government's budget deficit (or reduces its surplus) during a recession and increases government's budget surplus (or reduces its deficit) during an expansion without any action by policymakers. The tax system is one such mechanism.

Bureau of Economic Analysis (BEA) An agency of the U.S. Department of Commerce that compiles the national income and product accounts.

business cycle Recurring increases and decreases in the level of economic activity over periods of years; consists of peak, recession, trough, and expansion phases.

businesses Economic entities (*firms*) that purchase resources and provide *goods* and *services* to the economy.

business firm (See *firm*.)

cap-and-trade program A government strategy for reducing harmful emissions or discharges by placing a limit on their total amounts and then allowing *firms* to buy and sell the rights to emit or discharge specific amounts within the total limits.

capital Human-made resources (buildings, machinery, and equipment) used to produce *goods* and *services;* goods that do not directly satisfy human wants; also called capital goods. One of the four *economic resources*.

capital and financial account The section of a nation's *international balance of payments* that records (1) debt forgiveness by and to foreigners and (2) foreign purchases of assets in the United States and U.S. purchases of assets abroad.

capital and financial account deficit A negative balance on its *capital and financial account* in a country's *international balance of payments*.

capital and financial account surplus A positive balance on its *capital and financial account* in a country's *international balance of payments*.

capital flight (Web chapter) The transfer of savings from *developing countries* to *industrially advanced countries* to avoid government expropriation, taxation, or higher rates of *inflation*, or simply to realize greater returns on *financial investments*.

capital gain The gain realized when *securities* or properties or other assets are sold for a *price* greater than the price paid for them.

capital goods (See *capital*.)

capital-intensive goods Products that require relatively large amounts of *capital* to produce.

capitalism An economic system in which property resources are privately owned and markets and prices are used to direct and coordinate economic activities.

capital-saving technology (Web chapter) An improvement in *technology* that permits a greater quantity of a product to be produced with a specific amount of *capital* (or permits the same amount of the product to be produced with a smaller amount of capital).

capital stock The total available *capital* in a nation.

capital-using technology (Web chapter) An improvement in *technology* that requires the use of a greater amount of *capital* to produce a specific quantity of a product.

capricious-universe view (Web chapter) The view held by some people that fate and outside events, rather than hard work and enterprise, will determine their economic destinies.

cardinal utility Satisfaction (*utility*) that can be measured via cardinal numbers (1, 2, 3…), with all the mathematical properties of those numbers such as addition, subtraction, multiplication, and division being applicable.

cartel A formal agreement among *firms* (or countries) in an *industry* to set the *price* of a product and establish the outputs of the individual firms (or countries) or to divide the market for the product geographically.

causation A relationship in which the occurrence of one or more events brings about another event.

CEA (See *Council of Economic Advisers*.)

cease-and-desist order An order from a court or government agency to a corporation or individual to stop engaging in a specified practice.

ceiling price (See *price ceiling*.)

Celler-Kefauver Act The federal law of 1950 that amended the *Clayton Act* by prohibiting the acquisition of the assets of one *firm* by another firm when the effect would be less competition.

central bank A bank whose chief function is the control of the nation's *money supply;* in the United States, the Federal Reserve System.

central economic planning Government determination of the objectives of the economy and how resources will be directed to attain those goals.

***ceteris paribus* assumption** (See *other-things-equal assumption*.)

change in demand A movement of an entire *demand curve* or schedule such that the *quantity demanded* changes at every particular *price;* caused by a change in one or more of the *determinants of demand.*

change in quantity demanded A change in the *quantity demanded* along a fixed *demand curve* (or within a fixed demand schedule) as a result of a change in the *price* of the product.

change in quantity supplied A change in the *quantity supplied* along a fixed *supply curve* (or within a fixed supply schedule) as a result of a change in the product's *price*.

change in supply A movement of an entire *supply curve* or schedule such that the *quantity supplied* changes at every particular *price;* caused by a change in one or more of the *determinants of supply*.

Change to Win A loose federation of American unions that includes the Service Workers and Teamsters unions; the second largest union federation after the *AFL-CIO*.

checkable deposit Any deposit in a *commercial bank* or *thrift institution* against which a check may be written.

checkable-deposit multiplier (See *monetary multiplier*.)

check clearing The process by which funds are transferred from the checking accounts of the writers of checks to the checking accounts of the recipients of checks.

checking account A *checkable deposit* in a *commercial bank* or *thrift institution*.

circular flow diagram An illustration showing the flow of *resources* from *households* to *firms* and of products from firms to households. These flows are accompanied by reverse flows of money from firms to households and from households to firms.

Clayton Act The federal antitrust law of 1914 that strengthened the *Sherman Act* by making it illegal for *firms* to engage in certain specified practices including *tying contracts*, *interlocking directorates*, and certain forms of *price discrimination*.

closed economy An economy that neither exports nor imports *goods* and *services*.

closed shop A place of employment where only workers who are already members of a labor union may be hired.

Coase theorem The idea, first stated by economist Ronald Coase, that some *externalities* can be resolved through private negotiations among the affected parties.

cognitive biases Misperceptions or misunderstandings that cause *systematic errors*. Most result either (1) from *heuristics* that are prone to *systematic errors* or (2) because the brain is attempting to solve a type of problem (such as a calculus problem) for which it was not evolutionarily evolved and for which it has little innate capability.

coincidence of wants A situation in which the *good* or *service* that one trader desires to obtain is the same as that which another trader desires to give up and an item that the second trader wishes to acquire is the same as that which the first trader desires to surrender.

COLA (See *cost-of-living adjustment*.)

collective-action problem The idea that getting a group to pursue a common, collective goal gets harder the larger the group's size. Larger groups are more costly to organize and their members more difficult to motivate because the larger the group, the smaller each member's share of the benefits if the group succeeds.

collective bargaining The negotiation of labor contracts between *labor unions* and *firms* or government entities.

collective voice The function a *labor union* performs for its members as a group when it communicates their problems and grievances to management and presses management for a satisfactory resolution.

collusion A situation in which *firms* act together and in agreement (collude) to fix *prices*, divide a market, or otherwise restrict competition.

command system A method of organizing an economy in which property resources are publicly owned and government uses *central economic planning* to direct and coordinate economic activities; *socialism;* communism. Compare with *market system*.

commercial bank A *firm* that engages in the business of banking (accepts deposits, offers checking accounts, and makes loans).

commercial banking system All *commercial banks* and *thrift institutions* as a group.

communism (See *command system*.)

comparative advantage A situation in which a person or country can produce a specific product at a lower opportunity cost than some other person or country; the basis for specialization and trade.

compensating differences Differences in the *wages* received by workers in different jobs to compensate for the nonmonetary differences between the jobs.

compensating wage differential (See *compensating differences.*)

compensation to employees *Wages* and salaries plus wage and salary supplements paid by employers to workers.

competition The effort and striving between two or more independent rivals to secure the business of one or more third parties by offering the best possible terms.

competitive industry's short-run supply curve The horizontal summation of the short-run supply curves of the *firms* in a purely competitive *industry* (see *pure competition*); a curve that shows the total quantities collectively offered for sale at various *prices* by the firms in an industry in the short run.

competitive labor market A resource market in which a large number of (noncolluding) employers demand a particular type of labor supplied by a large number of nonunion workers.

complementary goods Products and *services* that are used together. When the *price* of one falls, the demand for the other increases (and conversely).

complementary resources Productive inputs that are used jointly with other inputs in the production process; resources for which a decrease in the *price* of one leads to an increase in the demand for the other.

compound interest The accumulation of money that builds over time in an investment or interest-bearing account as new interest is earned on previous interest that is not withdrawn.

concentration ratio The percentage of the total sales of an *industry* made by the four (or some other number) largest sellers in the industry.

conflict diamonds Diamonds that are mined and sold by combatants in war zones in Africa as a way to provide the currency needed to finance their military activities.

conglomerate merger The merger of two *firms* operating in separate industries or separate geographic areas so that neither firm is a supplier, customer, or competitor of the other; any merger that is neither a *horizontal merger* nor a *vertical merger*.

conglomerates *Firms* that produce *goods* and *services* in two or more separate industries.

constant-cost industry An *industry* in which the entry and exit of *firms* have no effect on the *prices* firms in the industry must pay for resources and thus no effect on production costs.

constant opportunity cost An *opportunity cost* that remains the same for each additional unit as a consumer (or society) shifts purchases (production) from one product to another along a straight-line *budget line* (*production possibilities curve*).

constant returns to scale The situation when a firm's *average total cost* of producing a product remains unchanged in the *long run* as the firm varies the size of its *plant* (and, hence, its output).

consumer equilibrium In marginal utility theory, the combination of goods purchased that maximizes *total utility* by applying the *utility-maximizing rule*. In indifference curve analysis, the combination of goods purchased that maximizes *total utility* by enabling the consumer to reach the highest *indifference curve*, given the consumer's *budget line* (or *budget constraint*).

consumer goods Products and *services* that satisfy human wants directly.

Consumer Price Index (CPI) An index that measures the *prices* of a fixed "market basket" of some 300 *goods* and *services* bought by a "typical" consumer.

consumer sovereignty The determination by consumers of the types and quantities of *goods* and *services* that will be produced with the scarce resources of the economy; consumers' direction of production through their *dollar votes*.

consumer surplus The difference between the maximum *price* a consumer is (or consumers are) willing to pay for an additional unit of a product and its market price; the triangular area below the demand curve and above the market price.

consumption of fixed capital An estimate of the amount of *capital* worn out or used up (consumed) in producing the *gross domestic product;* also called depreciation.

consumption schedule A table of numbers showing the amounts *households* plan to spend for *consumer goods* at different levels of *disposable income*.

contractionary fiscal policy A decrease in *government purchases* of *goods* and *services*, an increase in *net taxes*, or some combination of the two, for the purpose of decreasing *aggregate demand* and thus controlling *inflation*.

coordination failure A situation in which people do not reach a mutually beneficial outcome because they lack some way to jointly coordinate their actions; a possible cause of macroeconomic instability.

copayment The percentage of (say, health care) costs that an insured individual pays while the insurer pays the remainder.

copyright A legal protection provided to developers and publishers of books, computer software, videos, and musical compositions against the unauthorized copying of their works by others.

core inflation The underlying increases in the *price level* after volatile food and energy *prices* are removed.

corporate income tax A tax levied on the net income (accounting profit) of corporations.

corporation A legal entity ("person") chartered by a state or the federal government that is distinct and separate from the individuals who own it.

correlation A systematic and dependable association between two sets of data (two kinds of events); does not necessarily indicate causation.

corruption (Web chapter) The misuse of government power, with which one has been entrusted or assigned, to obtain private gain; includes payments from individuals or companies to secure advantages in obtaining government contracts, avoiding government regulations, or obtaining inside knowledge about forthcoming policy changes.

cost-benefit analysis A comparison of the *marginal costs* of a project or program with the *marginal benefits* to decide whether or not to employ resources in that project or program and to what extent.

cost-of-living adjustment (COLA) An automatic increase in the incomes (*wages*) of workers when *inflation* occurs; often included in *collective bargaining* agreements between *firms* and *unions*. Cost-of-living adjustments are also guaranteed by law for *Social Security* benefits and certain other government *transfer payments*.

cost-push inflation Increases in the *price level* (*inflation*) resulting from an increase in resource costs (for example, raw-material prices) and hence in *per-unit production costs;* inflation caused by reductions in *aggregate supply*.

Council of Economic Advisers (CEA) A group of three persons that advises and assists the president of the United States on economic matters (including the preparation of the annual *Economic Report of the President*).

counter-cyclical payments (CCPs) Cash *subsidies* paid to farmers when market *prices* for certain crops drop below targeted prices. Payments are based on previous production and are received regardless of the current crop grown.

craft union A labor union that limits its membership to workers with a particular skill (craft).

creative destruction The hypothesis that the creation of new products and production methods destroys the market power of existing monopolies.

credible threat In *game theory*, a statement of harmful intent by one party that the other party views as believable; often issued in conditional terms of "if you do this, we will do that."

credit An accounting entry that either increases the value of an *asset* or reduces the value of a *liability* (by acting as a deduction from an amount already owed).

credit union A financial institution that provides many of the same services as a *commercial bank*, including checkable deposits, but only to members, who own the credit union and who share a common tie (such as being employees of the same *firm* or members of the same labor union).

cross elasticity of demand The ratio of the percentage change in *quantity demanded* of one good to the percentage change in the *price* of some other good. A positive coefficient indicates the two products are *substitute goods;* a negative coefficient indicates they are *complementary goods*.

crowding model of occupational discrimination A model of labor markets suggesting that *occupational discrimination* has kept many women and minorities out of high-paying occupations and forced them into a limited number of low-paying occupations.

crowding-out effect A rise in interest rates and a resulting decrease in *planned investment* caused by the federal government's increased borrowing to finance budget deficits and refinance debt.

currency Coins and paper money.

currency appreciation (See *exchange-rate appreciation*.)

currency depreciation (See *exchange-rate depreciation*.)

currency intervention A government's buying and selling of its own currency or foreign currencies to alter international exchange rates.

current account The section in a nation's *international balance of payments* that records its exports and imports of *goods* and *services*, its net *investment income*, and its *net transfers*.

cyclical asymmetry The idea that *monetary policy* may be more successful in slowing expansions and controlling *inflation* than in extracting the economy from severe recession.

cyclical deficit a Federal *budget deficit* that is caused by a recession and the consequent decline in tax revenues.

cyclically adjusted budget The estimated annual budget deficit or surplus that would occur under existing tax rates and government spending levels if the the economy were to operate at its *full-employment* level of GDP for a year; the *full-employment* budget deficit or surplus.

cyclical unemployment A type of *unemployment* caused by insufficient total spending (insufficient *aggregate demand*) and which typically begins in the *recession* phase of the *business cycle*.

deadweight loss (See *efficiency loss*.)

debit An accounting entry that either decreases the value of an *asset* or increases the value of a *liability* (by increasing the size of an amount already owed).

debt crisis An economic crisis in which government debt has risen so high that the government is unable to borrow any more money due to people losing faith in the government's ability to repay. Leads to either massive spending cuts or large tax increases, either of which will likely plunge the economy into a *recession*.

declining industry An *industry* whose total output is declining because negative *economic profits* lead many *firms* to exit the industry.

decreasing-cost industry An *industry* in which expansion through the entry of *firms* lowers the *prices* that firms in the industry must pay for resources and therefore decreases their production costs.

deductible The dollar sum of (for example, health care) costs that an insured individual must pay before the insurer begins to pay.

defaults Situations in which borrowers stop making loan payments or do not pay back loans that they took out and are now due.

defensive medicine The recommendation by physicians of more tests and procedures than are warranted medically or economically as a way of protecting themselves against later malpractice suits.

deflating The process of using a *price index* to decrease (deflate) a given year's *nominal gross domestic product* down to the smaller value of its *real gross domestic product;* only applicable if the given year's *price level* is higher than the price level that prevailed during the price index's *base year*. Compare with *inflating*.

deflation A decline in the general level of *prices* in an economy; a decline in an economy's *price level*.

demand A schedule or curve that shows the various amounts of a product that consumers are willing and able to purchase at each of a series of possible *prices* during a specified period of time.

demand curve A curve that illustrates the *demand* for a product by showing how each possible *price* (on the *vertical axis*) is associated with a specific *quantity demanded* (on the *horizontal axis*).

demand factor (in growth) The requirement that *aggregate demand* increase as fast as *potential output* if *economic growth* is to proceed as quickly as possible.

demand management The use of *fiscal policy* and *monetary policy* to increase or decrease *aggregate demand*.

demand-pull inflation Increases in the *price level* (*inflation*) resulting from increases in *aggregate demand*.

demand schedule A table of numbers showing the amounts of a *good* or *service* buyers are willing and able to purchase at various *prices* over a specified period of time.

demand shocks Sudden, unexpected changes in demand.

demand-side market failures Underallocations of resources that occur when private demand curves understate consumers' full willingness to pay for a *good* or *service*.

demographers Scientists who study the characteristics of human populations.

demographic transition (Web chapter) The massive decline in birth rates that occurs once a developing country achieves higher standards of living because the perceived marginal cost of additional children begins to exceed the perceived marginal benefit.

dependent variable A variable that changes as a consequence of a change in some other (independent) variable; the "effect" or outcome.

depository institutions Firms that accept deposits of *money* from the public (businesses and persons); *commercial banks*, *savings and loan associations*, *mutual savings banks*, and *credit unions*.

depreciation (See *consumption of fixed capital*.)

depreciation (of the dollar) A decrease in the value of the dollar relative to another currency, so a dollar buys a smaller amount of the foreign currency and therefore of foreign goods.

deregulation The removal of most or even all of the government regulation and laws designed to supervise an industry. Sometimes undertaken to combat *regulatory capture*.

derived demand The demand for a resource that depends on the demand for the products it helps to produce.

determinants of aggregate demand Factors such as consumption spending, *investment*, government spending, and *net exports* that, if they change, shift the aggregate demand curve.

determinants of aggregate supply Factors such as input prices, *productivity*, and the legal-institutional environment that, if they change, shift the aggregate supply curve.

determinants of demand Factors other than *price* that determine the quantities demanded of a *good* or *service*. Also referred to as "demand shifters" because changes in the determinants of demand will cause the *demand curve* to shift either right or left.

determinants of supply Factors other than *price* that determine the quantities supplied of a *good* or *service*. Also referred to as "supply shifters" because changes in the determinants of supply will cause the *supply curve* to shift either right or left.

developing countries (Web chapter) Many countries of Africa, Asia, and Latin America that are characterized by lack of capital goods, use of nonadvanced technologies, low literacy rates, high unemployment, relatively rapid population growth, and labor forces heavily committed to agriculture.

diagnosis-related group (DRG) system Payments to doctors and hospitals under *Medicare* based on which of hundreds of carefully detailed diagnostic categories best characterize each patient's condition and needs.

dictator game A mutually anonymous behavioral economics game in which one person ("the dictator") unilaterally determines how to split an amount of money with the second player.

differentiated oligopoly An *oligopoly* in which *firms* produce a *differentiated product*.

differentiated product A product that differs physically or in some other way from the similar products produced by other *firms;* a product such that buyers are not indifferent to the seller when the *price* charged by all sellers is the same.

diffusion (Web chapter) The spread of an *innovation* through its widespread imitation.

dilemma of regulation The trade-off faced by a *regulatory agency* in setting the maximum legal *price* a monopolist may charge: The *socially optimal price* is below *average total cost* (and either bankrupts the *firm* or requires that it be subsidized), while the higher, *fair-return price* does not produce *allocative efficiency*.

diminishing marginal returns (See *law of diminishing returns*.)

diminishing marginal utility (See *law of diminishing marginal utility*.)

direct foreign investment (See *foreign direct investment*.)

direct payments Cash subsidies paid to farmers based on past production levels; a permanent transfer payment unaffected by current crop *prices* and current production.

direct relationship The relationship between two variables that change in the same direction, for example, product *price* and quantity supplied; a positive relationship.

discount rate The interest rate that the *Federal Reserve Banks* charge on the loans they make to *commercial banks* and *thrift institutions*.

discouraged workers Employees who have left the *labor force* because they have not been able to find employment.

discretionary fiscal policy Deliberate changes in taxes (tax rates) and government spending to promote full employment, price stability, and economic growth.

discrimination The practice of according individuals or groups inferior treatment in hiring, occupational access, education and training, promotion, wage rates, or working conditions even though they have the same abilities, education, skills, and work experience as other workers.

discrimination coefficient A measure of the cost or disutility of prejudice; the monetary amount an employer is willing to pay to hire a preferred worker rather than a nonpreferred worker of the same ability.

diseconomies of scale The situation when a firm's *average total cost* of producing a product increases in the *long run* as the firm increases the size of its *plant* (and, hence, its output).

disinflation A reduction in the rate of *inflation*.

disposable income (DI) *Personal income* less personal taxes; income available for *personal consumption expenditures* and *personal saving*.

dissaving Spending for *consumer goods* in excess of *disposable income;* the amount by which *personal consumption expenditures* exceed disposable income.

diversifiable risk Investment *risk* that investors can reduce via *diversification;* also called idiosyncratic risk.

diversification The strategy of investing in a large number of investments in order to reduce the overall risk to an entire investment *portfolio*.

dividends Payments by a corporation of all or part of its profit to its stockholders (the corporate owners).

division of labor The separation of the work required to produce a product into a number of different tasks that are performed by different workers; *specialization* of workers.

Doha Development Agenda The latest, uncompleted (as of late 2013) sequence of trade negotiations by members of the *World Trade Organization;* named after Doha, Qatar, where the set of negotiations began. Also called the Doha Round.

dollar votes The "votes" that consumers cast for the production of preferred products when they purchase those products rather than the alternatives that were also available.

domestic capital formation The process of adding to a nation's stock of *capital* by saving and investing part of *domestic output*.

domestic output *Gross* (or net) *domestic product;* the total output of *final goods* and final *services* produced in the economy.

domestic price The *price* of a *good* or *service* within a country, determined by domestic demand and supply.

dominant strategy In *game theory*, an option that is better for a player than any other alternative option regardless of what the player's opponent(s) may do.

dumping The sale of a product in a foreign country at *prices* either below cost or below the prices commonly charged at home.

DuPont cellophane case The antitrust case brought against DuPont in which the U.S. Supreme Court ruled (in 1956) that while DuPont had a monopoly in the narrowly defined market for cellophane, it did not monopolize the more broadly defined market for flexible packaging materials. It was thus not guilty of violating the *Sherman Act*.

durable good A consumer good with an expected life (use) of three or more years.

earmarks Narrow, specially designated spending authorizations placed in broad legislation by senators and representatives for the purpose of providing benefits to *firms* and organizations within their constituencies. Earmarked projects are exempt from competitive bidding and normal evaluation procedures.

earned-income tax credit (EITC) A refundable federal *tax credit* for low-income working people designed to reduce poverty and encourage labor-force participation.

earnings The money income received by a worker; equal to the *wage* (rate) multiplied by the amount of time worked.

economic concentration A description or measure of the degree to which an *industry* is dominated by one or a handful of *firms* or is characterized by many firms. (See *concentration ratio*.)

economic cost A payment that must be made to obtain and retain the *services* of a *resource;* the income a *firm* must provide to a resource supplier to attract the resource away from an alternative use; equal to the quantity of other products that cannot be produced when resources are instead used to make a particular product.

economic efficiency The use of the minimum necessary resources to obtain the socially optimal amounts of *goods* and *services;* entails both *productive efficiency* and *allocative efficiency*.

economic growth (1) An outward shift in the *production possibilities curve* that results from an increase in resource supplies or quality or an improvement in *technology;* (2) an increase of real output (*gross domestic product*) or real output per capita.

economic immigrants International migrants who have moved from one country to another to obtain economic gains such as better employment opportunities.

economic investment (See *investment.*)

economic law An *economic principle* that has stood the test of time.

economic model A simplified picture of economic reality; an abstract generalization.

economic perspective A viewpoint that envisions individuals and institutions making rational decisions by comparing the marginal benefits and marginal costs associated with their actions.

economic policy A course of action intended to correct or avoid a problem.

economic principle A widely accepted generalization about the economic behavior of individuals or institutions.

economic profit The return flowing to those who provide the economy with the *economic resource* of *entrepreneurial ability;* the *total revenue* of a *firm* less its *economic costs* (which include both *explicit costs* and *implicit costs*); also called "pure profit" and "above-normal profit."

economic regulation (See *industrial regulation* and *social regulation.*)

economic rent The *price* paid for the use of land and other natural resources that are in fixed (*perfectly inelastic*) supply.

economic resources The *land*, *labor*, *capital*, and *entrepreneurial ability* that are used to produce *goods* and *services;* the *factors of production*.

economics The social science concerned with how individuals, institutions, and society make optimal (best) choices under conditions of scarcity.

economic system A particular set of institutional arrangements and a coordinating mechanism for solving the *economizing* *problem*; a method of organizing an economy, of which the *market system* and the *command system* are the two general types.

economic theory A statement of a cause-effect relationship; when accepted by all or nearly all economists, an *economic principle*.

economies of scale The situation when a firm's *average total cost* of producing a product decreases in the *long run* as the firm increases the size of its *plant* (and, hence, its output).

economizing problem The choices necessitated because society's economic wants for *goods* and *services* are unlimited but the resources available to satisfy these wants are limited (scarce).

efficiency factor (in growth) The capacity of an economy to achieve *allocative efficiency* and *productive efficiency* and thereby fulfill the potential for growth that the *supply factors* (*of growth*) make possible; the capacity of an economy to achieve *economic efficiency* and thereby reach the optimal point on its *production possibilities curve*.

efficiency gains from migration The increases in total worldwide output that take place if the additions to output from *immigration* in the destination nation exceed the loss of output from *emigration* from the origin nation.

efficiency loss Reductions in combined consumer and producer surplus caused by an underallocation or overallocation of resources to the production of a *good* or *service*. Also called *deadweight loss*.

efficiency loss of a tax The loss of *net benefits* to society because a tax reduces the production and consumption of a taxed good below the level of *allocative efficiency*. Also called the *deadweight loss* of the tax.

efficiency wage An above-market (above-equilibrium) *wage* that minimizes wage costs per unit of output by encouraging greater effort or reducing turnover.

efficient allocation of resources That distribution of society's scarce *resources* that produces the socially optimal mix of output; *allocative efficiency*.

elastic demand Product or resource demand whose *price elasticity of demand* is greater than 1, so that any given percentage change in *price* leads to a larger percentage change in *quantity demanded*. As a result, quantity demanded is relatively sensitive to (elastic with respect to) price.

elasticity coefficient The number obtained when the percentage change in *quantity demanded* (or supplied) of a product or resource is divided by the percentage change in its *price*.

elasticity formula (See *price elasticity of demand*.)

elasticity of resource demand A measure of the responsiveness of *firms* to a change in the *price* of a particular *resource* they employ or use; the percentage change in the quantity demanded of the *resource* divided by the percentage change in its *price*.

elastic supply Product or resource supply whose *price elasticity of supply* is greater than 1, so that the resulting percentage change in *quantity supplied* is greater than the percentage change in *price*. As a result, quantity supplied is relatively sensitive to (elastic with respect to) price.

Electronic Benefit Transfer (EBT) cards Debit cards used by the federal government to deliver food money to low-income recipients as part of the *Supplemental Nutrition Assistance Program* (*SNAP*). The same cards are also used by some states to deliver the benefits issued by a variety of additional *public assistance programs*.

electronic payments Purchases made by transferring funds electronically. Examples include credit cards, debit cards, *Electronic Benefit Transfer* (*EBT*) *cards*, Fedwire transfers, automated clearinghouse transactions (ACHs), payments via the PayPal system, and payments made through stored-value cards.

emigration The exit (outflow) of residents from a country to reside in foreign countries.

employer mandate The requirement under the *Patient Protection and Affordable Care Act* (*PPACA*) of 2010 that firms with 50 or more employees pay for insurance policies for their employees or face a fine of $2,000 per employee per year. Firms with fewer than 50 employees are exempt.

employment rate The percentage of the *labor force* employed at any time.

empty threat In *game theory*, a statement of harmful intent that is easily dismissed by the recipient because the threat is not viewed as being believable; compare to *credible threat*.

endowment effect The tendency people have to place higher valuations on items they possess (are endowed with) than on identical items that they do not possess; perhaps caused by *loss aversion*.

entitlement programs Government programs such as *social insurance*, *Medicare*, and *Medicaid* that guarantee (entitle) particular levels of transfer payments or noncash benefits to all who fit the programs' criteria.

entrepreneurial ability The human resource that combines the other *economic resources* of *land*, *labor*, and *capital* to produce new products or make innovations in the production of existing products; provided by *entrepreneurs*.

entrepreneurs Individuals who provide *entrepreneurial ability* to *firms* by setting strategy, advancing innovations, and bearing the financial risk if their firms do poorly.

equality-efficiency trade-off The decrease in *economic efficiency* that may accompany a decrease in *income inequality;* the presumption that some income inequality is required to achieve economic efficiency.

equation of exchange $MV = PQ$, in which M is the supply of *money*, V is the *velocity* of money, P is the *price level*, and Q is the physical volume of *final goods* and final *services* produced.

equilibrium GDP (See *equilibrium real domestic output*.)

equilibrium position In the indifference curve model, the combination of two goods at which a consumer maximizes his or her *utility* (reaches the highest attainable *indifference curve*), given a limited amount to spend (a *budget constraint*).

equilibrium price The *price* in a competitive market at which the *quantity demanded* and the *quantity supplied* are equal, there is neither a shortage nor a surplus, and there is no tendency for price to rise or fall.

equilibrium price level In the *aggregate demand–aggregate supply* (*AD-AS*) *model*, the *price level* at which *aggregate demand* equals *aggregate supply;* the price level at which the aggregate demand curve intersects the aggregate supply curve.

equilibrium quantity (1) The quantity at which the intentions of buyers and sellers in a particular market match at a particular *price* such that the *quantity demanded* and the *quantity supplied* are equal; (2) the profit-maximizing output of a *firm*.

equilibrium real domestic output The *gross domestic product* at which the total quantity of *final goods* and final *services* purchased (*aggregate expenditures*) is equal to the total quantity of final goods services produced (the real domestic output); the real domestic output at which the aggregate demand curve intersects the aggregate supply curve.

equilibrium real output (See *equilibrium real domestic output*.)

equilibrium world price The *price* of an internationally traded product that equates the quantity of the product demanded by importers with the quantity of the product supplied by exporters; the price determined at the intersection of the export supply curve and the import demand curve.

euro The common *currency* unit used by 17 European nations (as of mid-2013) in the *eurozone*, which consists of Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovakia, Slovenia, and Spain.

European Union (EU) An association of 28 European nations (as of mid-2013) that has eliminated tariffs and quotas among them, established common tariffs for imported goods from outside the member nations, eliminated barriers to the free movement of capital, and created other common economic policies.

eurozone The 17 nations (as of 2013) of the 28-member (as of 2013) *European Union* that use the *euro* as their common *currency*. The eurozone countries are Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovakia, Slovenia, and Spain.

excess capacity *Plant* resources that are underused when imperfectly competitive *firms* produce less output than that associated with achieving minimum *average total cost*.

excess reserves The amount by which a *commercial bank*'s or *thrift institution*'s *actual reserves* exceed its *required reserves;* actual reserves minus required reserves.

exchange controls (See *foreign exchange controls*.)

exchange rate The *rate of exchange* of one nation's *currency* for another nation's currency.

exchange-rate appreciation An increase in the value of a nation's *currency* in *foreign exchange markets;* an increase in the *rate of exchange* with foreign currencies.

exchange-rate depreciation A decrease in the value of a nation's *currency* in *foreign exchange markets;* a decrease in the *rate of exchange* with foreign currencies.

exchange-rate determinant Any factor other than the *rate of exchange* that determines a *currency*'s demand and supply in the *foreign exchange market*.

excise tax A tax levied on the production of a specific product or on the quantity of the product purchased.

excludability The characteristic of a *private good*, for which the seller can keep nonbuyers from obtaining the good.

exclusive unionism The policy, pursued by many *craft unions*, in which a *union* first gets employers to agree to hire only union workers and then excludes many workers from joining the union so as to restrict the supply of labor and drive up wages. Compare with *inclusive unionism*. The policies typically employed by a *craft union*.

exhaustive expenditure An expenditure by government resulting directly in the employment of *economic resources* and in the absorption by government of the *goods* and *services* those resources produce; a *government purchase*.

exit mechanism The method of resolving workplace dissatisfaction by quitting one's job and searching for another.

expanding industry An *industry* whose total output is increasing because positive *economic profits* lead many *firms* to enter the industry.

expansion The phase of the *business cycle* in which *real GDP*, *income*, and employment rise.

expansionary fiscal policy An increase in *government purchases* of *goods* and *services*, a decrease in *net taxes*, or some combination of the two for the purpose of increasing *aggregate demand* and expanding real output.

expansionary monetary policy *Federal Reserve System* actions to increase the *money supply*, lower *interest rates*, and expand *real GDP;* an easy money policy.

expectations The anticipations of consumers, *firms*, and others about future economic conditions.

expected rate of return The increase in profit a *firm* anticipates it will obtain by purchasing capital or engaging in research and development (*R&D*); expressed as a percentage of the total cost of the investment (or R&D) activity.

expected-rate-of-return curve (Web chapter) As it relates to research and development (*R&D*), a curve showing the anticipated gain in *profit*, as a percentage of R&D expenditure, from an additional dollar spent on R&D.

expenditures approach The method that adds all expenditures made for *final goods* and final *services* to measure the *gross domestic product*.

expenditures-output approach (See *aggregate expenditures–domestic output approach*.)

explicit cost The monetary payment made by a *firm* to an outsider to obtain a *resource*.

exports *Goods* and *services* produced in a nation and sold to buyers in other nations.

export subsidy A government payment to a domestic producer to enable the *firm* to reduce the *price* of a *good* or *service* to foreign buyers.

export supply curve An upward-sloping curve that shows the amount of a product that domestic *firms* will export at each *world price* that is above the *domestic price*.

export transaction A sale of a *good* or *service* that increases the amount of foreign currency flowing to a nation's citizens, *firms*, and government.

external benefit (See *positive externality*.)

external cost (See *negative externality*.)

external debt Private or public debt owed to foreign citizens, firms, and institutions.

externality A cost or benefit from production or consumption that accrues to to someone other than the immediate buyers and sellers of the product being produced or consumed (see *negative externality* and *positive externality*).

external public debt The portion of the public debt owed to foreign citizens, *firms*, and institutions.

extraction cost All costs associated with extracting a natural resource and readying it for sale.

face value The numerical value printed or inscribed on coins or paper money.

factors of production The four *economic resources: land*, *labor*, *capital*, and *entrepreneurial ability*.

fairness A person's opinion as to whether a price, wage, or allocation is considered morally or ethically acceptable.

fair-return price For *natural monopolies* subject to rate (*price*) regulation, the price that would allow the regulated monopoly to earn a *normal profit;* a price equal to *average total cost*.

fallacy of composition The false notion that what is true for the individual (or part) is necessarily true for the group (or whole).

farm commodities Agricultural products such as grains, milk, cattle, fruits, and vegetables that are usually sold to processors, who use the products as inputs in creating *food products*.

fast-second strategy (Web chapter) An approach by a dominant *firm* in which it allows other firms in its *industry* to bear the risk of innovation and then quickly becomes the second firm to offer any successful new product or adopt any improved production process.

FDIC (See *Federal Deposit Insurance Corporation*.)

Federal Deposit Insurance Corporation (FDIC) The federally chartered corporation that insures the deposit liabilities (up to $250,000 per account) of *commercial banks* and *thrift institutions* (excluding *credit unions*, whose deposits are insured by the *National Credit Union Administration*).

federal funds rate The interest rate that U.S. banks and other depository institutions charge one another on overnight loans made out of their *excess reserves*.

federal government The government of the United States, as distinct from the state and local governments.

Federal Open Market Committee (FOMC) The 12-member group within the *Federal Reserve System* that decides U.S. *monetary policy* and how it is executed through *open-market operations* (in which the Fed buys and sells U.S. government securities to adjust the *money supply*.)

Federal Reserve Banks The 12 banks chartered by the U.S. government that collectively act as the *central bank* of the United States. They set monetary policy and regulate the private banking system under the direction of the *Board of Governors* and the *Federal Open Market Committee*. Each of the 12 is a *quasi-public bank* and acts as a *banker's bank* in its designated geographic region.

Federal Reserve Note Paper money issued by the *Federal Reserve Banks*.

Federal Reserve System The U.S. central bank, consisting of the *Board of Governors* of the Federal Reserve and the 12 *Federal Reserve Banks*, which controls the lending activity of the nation's banks and thrifts and thus the *money supply;* commonly referred to as the "Fed."

Federal Trade Commission (FTC) The commission of five members established by the *Federal Trade Commission Act* of 1914 to investigate unfair competitive practices of firms, to hold hearings on the complaints of such practices, and to issue *cease-and-desist orders* when *firms* are found to engage in such practices.

Federal Trade Commission Act The federal law of 1914 that established the *Federal Trade Commission*.

fee for service In the health care *industry*, payment to physicians for each visit made or procedure performed.

fiat money Anything that is *money* because government has decreed it to be money.

final goods Goods that have been purchased for final use (rather than for resale or further processing or manufacturing.)

financial capital (See *money capital*.)

financial investment The purchase of a financial asset (such as a *stock*, *bond*, or *mutual fund*) or real asset (such as a house, land, or factories) or the building of such assets in the expectation of financial gain.

financial services industry The broad category of *firms* that provide financial products and services to help *households* and businesses earn *interest*, receive *dividends*, obtain *capital gains*, insure against losses, and plan for retirement. Includes *commercial banks*, *thrift institutions*, insurance companies, mutual fund companies, pension funds, investment banks, and *securities* firms.

firm An organization that employs resources to produce a *good* or *service* for profit and owns and operates one or more *plants*.

first-mover advantage In *game theory*, the benefit obtained by the party that moves first in a *sequential game*.

fiscal policy Changes in government spending and tax collections designed to achieve full employment, price stability, and economic growth; also called *discretionary fiscal policy*.

fishery A stock of fish or other marine animal that is composed of a distinct group, for example New England cod, Pacific tuna, or Alaskan crab.

fishery collapse A rapid decline in a *fishery*'s population because its fish are being harvested faster than they can reproduce.

fixed cost Any cost that in total does not change when the *firm* changes its output.

fixed exchange rate A *rate of exchange* that is set in some way and therefore prevented from rising or falling with changes in currency supply and demand.

fixed resource Any resource whose quantity cannot be changed by a firm in the *short run*.

flexible exchange rate A *rate of exchange* that is determined by the international demand for and supply of a nation's money and that is consequently free to rise or fall because it is not subject to *currency interventions*. Also referred to as a "floating exchange rate."

flexible prices Product *prices* that freely move upward or downward when product demand or supply changes.

floating exchange rate (See *flexible exchange rate*.)

follower countries As it relates to *economic growth*, countries that adopt advanced technologies that previously were developed and used by *leader countries*.

Food, Conservation, and Energy Act of 2008 Farm legislation that continued and extended three types of agricultural subsidies: *direct payments*, *counter-cyclical payments*, and the *marketing loan program*.

food products Processed *farm commodities* sold through grocery stores and restaurants. Examples: bread, meat, fish, chicken, pork, lettuce, peanut butter, and breakfast cereal.

foreign competition (See *import competition*.)

foreign direct investment (Web chapter) *Financial investments* made to obtain a lasting ownership interest in *firms* operating outside the economy of the investor; may involve purchasing existing assets or building new production facilities.

foreign exchange controls Restrictions that a government may impose over the quantity of foreign currency demanded by its citizens and *firms* and over the *rate of exchange* as a way to limit the nation's quantity of *outpayments* relative to its quantity of *inpayments* (in order to eliminate a *payments deficit*).

foreign exchange market A market in which the money (currency) of one nation can be used to purchase (can be exchanged for) the money of another nation; a currency market.

foreign exchange rate (See *rate of exchange*.)

foreign purchases effect The inverse relationship between the *net exports* of an economy and its *price level* relative to foreign price levels.

45° (degree) line The reference line in a two-dimensional graph that shows equality between the variable measured on the *horizontal axis* and the variable measured on the *vertical axis*. In the *aggregate expenditures model*, the line along which the value of output (measured horizontally) is equal to the value of *aggregate expenditures* (measured vertically).

forward commitment A policy statement by a *central bank* indicating that it will continue to pursue a *monetary policy* action until a certain date is reached or until some particular threshold has been reached (for instance, the unemployment rate falling below, say, seven percent).

four-firm concentration ratio The percentage of total *industry* sales accounted for by the top four *firms* in an industry.

fractional reserve banking system A system in which *commercial banks* and *thrift institutions* hold less than 100 percent of their checkable-deposit liabilities as reserves of *currency* held in bank vaults or as deposits at the *central bank*.

framing effects In *prospect theory*, changes in people's decision making caused by new information that alters the context, or "frame of reference," that they use to judge whether options are viewed as gains or losses relative to the *status quo*.

freedom of choice The freedom of owners of property resources to employ or dispose of them as they see fit, of workers to enter any line of work for which they are qualified, and of consumers to spend their incomes in a manner that they think is appropriate.

freedom of enterprise The freedom of *firms* to obtain economic resources, to use those resources to produce products of the firm's own choosing, and to sell their products in markets of their choice.

Freedom to Farm Act A law passed in 1996 that revamped 60 years of U.S. farm policy by ending *price supports* and *acreage allotments* for wheat, corn, barley, oats, sorghum, rye, cotton, and rice.

free-rider problem The inability of potential providers of an economically desirable *good* or *service* to obtain payment from those who benefit, because of *nonexcludability*.

free trade The absence of artificial (government-imposed) barriers to trade among individuals and *firms* in different nations.

frictional unemployment A type of unemployment caused by workers voluntarily changing jobs and by temporary layoffs; unemployed workers between jobs.

fringe benefits The forms of compensation other than *wages* that employees receive from their employers. Includes pensions, medical and dental insurance, paid vacation, and sick leave.

full employment (1) The use of all available resources to produce want-satisfying *goods* and *services;* (2) the situation in which the *unemployment rate* is equal to the *full-employment rate of unemployment;* there exist *frictional unemployment* and *structural unemployment* but not *cyclical unemployment;* and *real GDP* equals *potential output*.

full-employment rate of unemployment The *unemployment rate* at which there is no *cyclical unemployment* of the *labor force;* equal to between 5 and 6 percent in the United States because some *frictional* and *structural unemployment* are unavoidable.

functional distribution of income The manner in which *national income* is divided among the functions performed to earn it (or the kinds of resources provided to earn it); the division of national income into wages and salaries, proprietors' income, corporate profits, interest, and rent.

future value The amount to which some current amount of *money* will grow if *interest* earned on the amount is left to compound over time. (*See compound interest.*)

G8 nations A group of eight major nations (Canada, France, Germany, Italy, Japan, Russia, United Kingdom, and United States) whose leaders meet regularly to discuss common economic problems and ways in which they might coordinate economic policies.

gains from trade The extra output that trading partners obtain through specialization of production and exchange of *goods* and *services*.

game theory The study of how people behave in strategic situations in which individuals must take into account not only their own possible actions but also the possible reactions of others. Originally developed to analyze the best ways to play games like poker and chess.

GDP (See *gross domestic product*.)

GDP gap Actual *gross domestic product* minus *potential output*; may be either a positive amount (a *positive GDP gap*) or a negative amount (a *negative GDP gap*).

GDP price index A *price index* for all the *goods* and *services* that make up the *gross domestic product;* the price index used to adjust *nominal gross domestic product* to *real gross domestic product*.

General Agreement on Tariffs and Trade (GATT) The international agreement reached in 1947 in which 23 nations agreed to eliminate *import quotas*, negotiate reductions in *tariff* rates, and give each other equal and nondiscriminatory treatment. It now includes most nations and has become the *World Trade Organization*.

Gini ratio A numerical measure of the overall dispersion of income among *households*, families, or individuals; found graphically by dividing the area between the diagonal line and the *Lorenz curve* by the entire area below the diagonal line.

gold standard A historical system of fixed exchange rates in which nations defined their currencies in terms of gold, maintained fixed relationships between their stocks of gold and their money supplies, and allowed gold to be freely exported and imported.

good Merchandise; an article of trade; a manufactured item offered for sale to consumers.

government failure Inefficiencies in resource allocation caused by problems in the operation of the *public sector* (government). Specific examples include the *principal-agent problem*, the *special-interest effect*, the *collective-action problem*, *rent seeking*, and *political corruption*.

government purchases (*G*) Expenditures by government for *goods* and *services* that government consumes in providing public services as well as expenditures for publicly owned capital that has a long lifetime; the expenditures of all governments in the economy for those *final goods* and final *services*.

government transfer payment Any disbursement of money by the government for which the government receives no currently produced *good* or *service* in return. Includes payments made by *public assistance programs* ("welfare") and *Social Security*.

grievance procedure The method used by a *labor union* and a *firm* to settle disputes that arise during the life of the *collective bargaining* agreement between them.

gross domestic product (GDP) The total market value of all *final goods* and final *services* produced annually within the boundaries of a nation.

gross private domestic investment (I_g) Expenditures for newly produced *capital goods* (such as machinery, equipment, tools, and buildings) and for additions to inventories.

growth accounting The bookkeeping of the supply-side elements such as productivity and labor inputs that contribute to changes in *real GDP* over some specific time period.

guiding function of prices The ability of *price* changes to bring about changes in the quantities of products and resources demanded and supplied.

H1-B provision A provision of the U.S. immigration law that allows the annual entry of 65,000 high-skilled workers in "specialty occupations" such as science, *R&D*, and computer programming to work legally and continuously in the United States for six years.

health maintenance organizations (HMOs) Health care providers that contract with employers, insurance companies, labor unions, or government units to provide health care for their workers or others who are insured.

health savings accounts (HSAs) Tax-free savings accounts into which people with high-deductible health insurance plans can place funds each year. Accumulated funds can be used to pay out-of-pocket medical expenses such as *deductibles* and *copayments*. Unused funds accumulate from year to year and can be used after retirement to supplement *Medicare*.

Herfindahl index A measure of the concentration and competitiveness of an *industry;* calculated as the sum of the squared percentage market shares of the individual *firms* in the industry.

heuristics The brain's low-energy mental shortcuts for making decisions. They are "fast and frugal" and work well in most situations but in other situations result in *systematic errors*.

homogeneous oligopoly An *oligopoly* in which *firms* produce a *standardized product*.

horizontal axis The "left-right" or "west-east" measurement line on a graph or grid.

horizontal merger The merger into a single *firm* of two firms producing the same product and selling it in the same geographic market.

households Economic entities (of one or more persons occupying a housing unit) that provide *resources* to the economy and use the *income* received to purchase *goods* and *services* that satisfy economic wants.

human capital The knowledge and skills that make a person productive.

human capital investment Any expenditure to improve the education, skills, health, or mobility of workers; normally undertaken with an expectation of greater productivity and thus a positive return on the investment.

hyperinflation A very rapid rise in the *price level;* an extremely high rate of *inflation*.

hypothesis A tentative explanation of cause and effect that requires testing to determine whether or not it is true.

illegal immigrants People who have entered a country unlawfully to reside there; also called unauthorized immigrants.

IMF (See *International Monetary Fund*.)

imitation problem (Web chapter) The potential for a *firm*'s rivals to produce a close variation of (imitate) a firm's new product or process, greatly reducing the originator's profit from *R&D* and *innovation*.

immediate market period The length of time during which the producers of a product are unable to change the quantity supplied in response to a change in price and in which there is a *perfectly inelastic supply*.

immediate short-run aggregate supply curve An *aggregate supply* curve for which real *output*, but not the *price level*, changes when the *aggregate demand* curve shifts; a horizontal *aggregate supply* curve that implies an inflexible price level.

immigration The inflow of people into a country from another country. The immigrants may be either *legal immigrants* or *illegal immigrants.*

immobility The inability or unwillingness of a worker to move from one geographic area or occupation to another or from a lower-paying job to a higher-paying job.

imperfect competition All *market structures* except *pure competition;* includes *monopoly*, *monopolistic competition*, and *oligopoly*.

implicit cost The monetary income a *firm* sacrifices when it uses a *resource* it owns rather than supplying the resource in the market; equal to what the resource could have earned in the best-paying alternative employment; includes a *normal profit*.

import competition The competition that domestic *firms* encounter from the products and *services* of foreign producers.

import demand curve A downsloping curve showing the amount of a product that an economy will import at each *world price* below the *domestic price*.

import quota A limit imposed by a nation on the quantity (or total value) of a good that may be imported during some period of time.

imports Spending by individuals, *firms*, and governments for *goods* and *services* produced in foreign nations.

import transaction The purchase of a *good* or *service* that decreases the amount of foreign money held by the citizens, *firms*, or government of a nation.

incentive function The inducement that an increase in the price of a commodity gives to sellers to make more of it available (and conversely for a decrease in *price*), and the inducement that an increase in price offers to buyers to purchase smaller quantities (and conversely for a decrease in price).

incentive pay plan A compensation structure that ties worker pay directly to performance. Such plans include piece rates, bonuses, *stock options*, commissions, and *profit-sharing plans.*

inclusive unionism The policy, pursued by *industrial unions*, in which a *union* attempts to include every worker in a given *industry* so as to be able to restrict the entire industry's labor supply and thereby raise wages. Compare with *exclusive unionism*.

income A flow of dollars (or *purchasing power*) per unit of time derived from the use of human or property resources.

income approach The method that adds all the *income* generated by the production of *final goods* and final *services* to measure the *gross domestic product*.

income effect A change in the quantity demanded of a product that results from the change in *real income* (*purchasing power*) caused by a change in the product's *price*.

income elasticity of demand The ratio of the percentage change in the *quantity demanded* of a good to a percentage change in consumer *income;* measures the responsiveness of consumer purchases to income changes.

income inequality The unequal distribution of an economy's total *income* among *households* or families.

income-maintenance system A group of government programs designed to eliminate poverty and reduce inequality in the distribution of income.

income mobility The extent to which *income* receivers move from one part of the income distribution to another over some period of time.

increase in demand An increase in the *quantity demanded* of a *good* or *service* at every price; a shift of the *demand curve* to the right. Caused by a change in one or more of the *determinants of demand*.

increase in supply An increase in the *quantity supplied* of a *good* or *service* at every price; a shift of the *supply curve* to the right. Caused by a change in one or more of the *determinants of supply*.

increasing-cost industry An *industry* in which expansion through the entry of new *firms* raises the *prices* firms in the industry must pay for *resources* and therefore increases their production costs.

increasing marginal returns An increase in the *marginal product* of a resource as successive units of the resource are employed.

increasing returns An increase in a *firm*'s output by a larger percentage than the percentage increase in its inputs.

independent goods Products or *services* for which there is little or no relationship between the *price* of one and the *demand* for the other. When the price of one rises or falls, the demand for the other tends to remain constant.

independent unions U.S. unions that are not affiliated with the *AFL-CIO* or *Change to Win*.

independent variable The variable causing a change in some other (dependent) variable.

index funds *Mutual funds* whose *portfolios* exactly match a stock or bond index (a collection of *stocks* or *bonds* meant to capture the overall behavior of a particular category of investments) such as the Standard & Poor's 500 Index or the Russell 3000 Index.

indifference curve A curve showing the different combinations of two products that yield the same satisfaction or *utility* to a consumer.

indifference map A set of *indifference curves*, each representing a different level of *utility*, that together show the preferences of a consumer.

individual demand The demand schedule or *demand curve* of a single buyer.

individual supply The supply schedule or *supply curve* of a single seller.

individual transferable quotas (ITQs) Limits (quotas) set by a government or a fisheries commission on the total number or total weight of a species that an individual fisher can harvest during some particular time period; fishers can sell (transfer) the right to use all or part of their respective individual quotas to other fishers.

industrially advanced countries (Web chapter) High-income countries such as the United States, Canada, Japan, and the nations of western Europe that have highly developed *market economies* based on large stocks of technologically advanced *capital goods* and skilled labor forces.

industrial regulation The older and more traditional type of regulation in which government is concerned with the *prices* charged and the *services* provided to the public in specific *industries*. Differs from *social regulation*.

industrial union A *labor union* that accepts as members all workers employed in a particular *industry* (or by a particular *firm*).

industry A group of (one or more) *firms* that produce identical or similar products.

inelastic demand Product or resource demand for which the *price elasticity of demand* is less than 1, so that any given percentage change in *price* leads to a smaller percentage change in *quantity demanded*. As a result, quantity demanded is relatively insensitive to (inelastic with respect to) price.

inelastic supply Product or resource supply for which the *price elasticity of supply* is less than 1, so that the resulting percentage change in *quantity supplied* is smaller than the percentage change in *price*. As a result, quantity supplied is relatively insensitive to (inelastic with respect to) price.

inferior good A *good* or *service* whose consumption declines as *income* rises, *prices* held constant.

inflating The process of using a *price index* to increase (inflate) a given year's *nominal gross domestic product* up to the larger value of its *real gross domestic product;* only applicable if the given year's *price level* is lower than the price level that prevailed during the price index's *base year*. Compare with *deflating*.

inflation A rise in the general level of *prices* in an economy; an increase in an economy's *price level*.

inflationary expectations The belief of workers, *firms*, and consumers about future rates of *inflation*.

inflationary expenditure gap In the *aggregate-expenditures model*, the amount by which the *aggregate expenditures schedule* must shift downward to decrease the *nominal GDP* to its full-employment noninflationary level.

inflation premium The component of the *nominal interest rate* that reflects anticipated *inflation*.

inflation targeting The annual statement by a *central bank* of a goal for a specific range of *inflation* in a future year, coupled with *monetary policy* designed to achieve the goal.

inflexible prices Product *prices* that remain in place (at least for a while) even though *supply* or *demand* has changed; stuck prices or sticky prices.

information technology New and more efficient methods of delivering and receiving information through the use of computers, Wi-Fi networks, wireless phones, and the Internet.

infrastructure The interconnected network of large-scale *capital goods* (such as roads, sewers, electrical grids, railways, ports, and the Internet) needed to operate a technologically advanced economy.

injection An addition of spending into the income-expenditure stream: any increment to *consumption*, *investment*, *government purchases*, or *net exports*.

injunction A court order directing a person or organization not to perform a certain act because the act would do irreparable damage to some other person or persons; a restraining order.

in-kind transfer The distribution by government of *goods* and *services* to individuals for which the government receives no currently produced *good* or *service* in return. Also called a noncash transfer because ordinary *transfer payments* are transfers of *money* rather than goods and services.

innovation (Web chapter) The first commercially successful introduction of a new product, use of a new method of production, or creation of a new form of business organization.

inpayments The receipts of domestic or foreign money that individuals, *firms*, and governments of one nation obtain from the sale of *goods* and *services* abroad, as investment income and remittances, and from foreign purchases of domestic assets.

insider-outsider theory The hypothesis that nominal wages are inflexible downward because *firms* are aware that workers ("insiders") who retain employment during recession may refuse to work cooperatively with previously unemployed workers ("outsiders") who offer to work for less than the current wage.

insurable risk An eventuality for which both the frequency and magnitude of potential losses can be estimated with considerable accuracy. Insurance companies are willing to sell insurance against such risks.

insurance exchanges Government-regulated markets for health insurance in which individuals seeking to purchase health insurance to comply with the *personal mandate* of the *Patient Protection and Affordable Care Act* (*PPACA*) of 2010 will be able to comparison shop among insurance policies approved by regulators. Each state will have its own exchange.

interest The payment made for the use of (borrowed) *money*.

interest income Payments of income to those who supply the economy with *capital*.

interest on reserves The payment by a *central bank* of *interest* on the deposits (*required reserves* plus *excess reserves*, if any) held by *commercial banks* at the central bank.

interest rate The annual rate at which *interest* is paid; a percentage of the borrowed amount.

interest-rate-cost-of-funds curve (Web chapter) As it relates to research and development (*R&D*), a curve showing the *interest rate* a *firm* must pay to obtain any particular amount of funds to finance R&D.

interest-rate effect The tendency for increases in the *price level* to increase the demand for money, raise interest rates, and, as a result, reduce total spending and real output in the economy (and the reverse for *price-level* decreases).

interindustry competition The competition for sales between the products of one *industry* and the products of another industry.

interlocking directorate A situation where one or more members of the board of directors of a *corporation* are also on the board of directors of a competing corporation; illegal under the *Clayton Act*.

intermediate goods Products that are purchased for resale or further processing or manufacturing.

internally held public debt *Public debt* owed to citizens, *firms*, and institutions of the same nation that issued the debt.

international balance of payments (See *balance of payments*.)

international balance-of-payments deficit (See *balance-of-payments deficit*.)

international balance-of-payments surplus (See *balance-of-payments surplus*.)

international gold standard (See *gold standard*.)

International Monetary Fund (IMF) The international association of nations that was formed after the Second World War to make loans of foreign monies to nations with temporary *balance of payments deficits* and, until the early 1970s, manage the international system of pegged exchange rates agreed upon at the Bretton Woods conference. It now mainly makes loans to nations facing possible defaults on private and government loans.

international monetary reserves The foreign currencies and other assets such as gold that a nation can use to settle a *balance-of-payments deficit*.

international value of the dollar The *price* that must be paid in foreign currency (money) to obtain one U.S. dollar.

intertemporal choice A choice between the benefits obtainable in one time period and the benefits obtainable in a later time period; the comparisons that individuals and society must make between the reductions in current consumption that are necessary to fund current investments and the higher levels of future consumption that those current investments will produce.

intrinsic value The market value of the metal within a coin.

invention (Web chapter) The conception of a new product or process combined with the first proof that it will work.

inventories Goods that have been produced but remain unsold.

inverse relationship The relationship between two variables that change in opposite directions, for example, product *price* and quantity demanded; a negative relationship.

inverted-U theory (Web chapter) The idea that, other things equal, *R&D* expenditures as a percentage of sales rise with *industry* concentration, reach a peak at a *four-firm concentration ratio* of about 50 percent, and then fall as the ratio further increases.

investment In economics, spending for the production and accumulation of *capital* and additions to *inventories*. (For contrast, see *financial investment*.)

investment banks *Firms* that help corporations and government raise money by selling *stocks* and *bonds;* they also offer advisory services for corporate mergers and acquisitions in addition to providing brokerage services and financial advice.

investment demand curve A curve that shows the amounts of *investment* demanded by an economy at a series of *real interest rates*.

investment goods Same as *capital* and *capital goods*.

investment in human capital (See *human capital investment*.)

investment schedule A table of numbers that shows the amounts *firms* plan to invest at various possible values of *real gross domestic product*.

"invisible hand" The tendency of *competition* to cause individuals and *firms* to unintentionally but quite effectively promote the interests of society even when each individual or firm is only attempting to pursue its own interests.

Joint Economic Committee (JEC) Committee of senators and representatives that investigates economic problems of national interest.

kinked-demand curve A *demand curve* that has a flatter slope above the current *price* than below the current price. Applies to a *noncollusive oligopoly* firm if its rivals will match any price decrease but ignore any price increase.

labor Any mental or physical exertion on the part of a human being that is used in the production of a *good* or *service*. One of the four *economic resources*.

labor force Persons 16 years of age and older who are not in institutions and who are employed or are unemployed and seeking work.

labor-force participation rate The percentage of the working-age population that is actually in the *labor force*.

labor-intensive goods Products requiring relatively large amounts of *labor* to produce.

labor productivity Total output divided by the quantity of labor employed to produce it; the *average product* of labor or output per hour of work.

labor union A group of workers organized to advance the interests of the group (to increase wages, shorten the hours worked, improve working conditions, and so on).

Laffer Curve A curve relating government tax rates and tax revenues and on which a particular tax rate (between zero and 100 percent) maximizes tax revenues.

laissez-faire capitalism A hypothetical *economic system* in which the government's economic role is limited to protecting private property and establishing a legal environment appropriate to the operation of *markets* in which only mutually agreeable transactions would take place between buyers and sellers; sometimes referred to as "pure *capitalism*."

land In addition to the part of the earth's surface not covered by water, this term refers to any and all natural resources ("free gifts of nature") that are used to produce *goods* and *services*. Thus, it includes the oceans, sunshine, coal deposits, forests, the electromagnetic spectrum, and *fisheries*. Note that land is one of the four *economic resources*.

land-intensive goods Products requiring relatively large amounts of land to produce.

land reform (Web chapter) Policy changes aimed at creating a more efficient distribution of land ownership in developing countries. Can involve everything from government purchasing large land estates and dividing the land into smaller farms to consolidating tiny plots of land into larger, more efficient private farms.

law of demand The principle that, other things equal, an increase in a product's *price* will reduce the quantity of it demanded, and conversely for a decrease in price.

law of diminishing marginal utility The principle that as a consumer increases the consumption of a *good* or *service*, the *marginal utility* obtained from each additional unit of the good or service decreases.

law of diminishing returns The principle that as successive increments of a variable *resource* are added to a fixed resource, the *marginal product* of the variable resource will eventually decrease.

law of increasing opportunity costs The principle that as the production of a good increases, the *opportunity cost* of producing an additional unit rises.

law of supply The principle that, other things equal, an increase in the *price* of a product will increase the quantity of it supplied, and conversely for a price decrease.

leader countries As it relates to *economic growth*, countries that develop and use the most advanced technologies, which then become available to *follower countries.*

leakage (1) A withdrawal of potential spending from the income-expenditures stream via *saving*, tax payments, or *imports;* (2) a withdrawal that reduces the lending potential of the banking system.

learning by doing Achieving greater *productivity* and lower *average total cost* through gains in knowledge and skill that accompany repetition of a task; a source of *economies of scale*.

least-cost combination of resources The quantity of each *resource* that a *firm* must employ in order to produce a particular output at the lowest total cost; the combination at which the ratio of the *marginal product* of a resource to its *marginal resource cost* (to its *price* if the resource is employed in a competitive market) is the same for the last dollar spent on each of the resources employed.

legal cartel theory of regulation The hypothesis that some *industries* seek regulation or want to maintain regulation so that they may form or maintain a legal *cartel*.

legal immigrant A person who lawfully enters a country for the purpose of residing there.

legal tender Any form of *currency* that by law must be accepted by creditors (lenders) for the settlement of a financial debt; a nation's official currency is legal tender within its own borders.

liability A debt with a monetary value; an amount owed by a *firm* or an individual.

limited liability rule A law that limits the potential losses that an investor in a *corporation* may suffer to the amount that she paid for her shares in the corporation. Encourages *financial investment* by limiting risk.

liquidity The degree to which an asset can be converted quickly into cash with little or no loss of purchasing power. *Money* is said to be perfectly liquid, whereas other assets have lesser degrees of liquidity.

liquidity trap A situation in a severe *recession* in which the *central bank's* injection of additional reserves into the banking system has little or no additional positive impact on lending, borrowing, *investment*, or *aggregate demand*.

loanable funds *Money* available for lending and borrowing.

loanable funds theory of interest The concept that the supply of and demand for *loanable funds* determine the equilibrium rate of *interest*.

loan guarantees A type of investment *subsidy* in which the government agrees to guarantee (pay off) the money borrowed by a private company to fund investment projects if the private company itself fails to repay the loan.

lockout A negotiating tactic in which a *firm* forbids its unionized workers to return to work until a new *collective bargaining* agreement is signed; a means of imposing costs (lost wages) on union workers.

logrolling The trading of votes by legislators to secure favorable outcomes on decisions concerning the provision of *public goods* and *quasi-public goods*.

long run (1) In *microeconomics*, a period of time long enough to enable producers of a product to change the quantities of all the resources they employ, so that all resources and costs are variable and no resources or costs are fixed. (2) In *macroeconomics*, a period sufficiently long for *nominal wages* and other input *prices* to change in response to a change in a nation's *price level*.

long-run aggregate supply curve The *aggregate supply* curve associated with a time period in which input *prices* (especially *nominal wages*) are fully responsive to changes in the *price level*.

long-run competitive equilibrium The *price* at which *firms* in *pure competition* neither obtain *economic profit* nor suffer economic losses in the *long run* and in which the total quantity demanded and supplied are equal; a price equal to the *marginal cost* and the minimum long-run *average total cost* of producing the product.

long-run supply In *microeconomics*, a schedule or curve showing the *prices* at which a purely competitive *industry* will make various quantities of the product available in the *long run*.

long-run supply curve As it applies to *macroeconomics*, a *supply curve* for which *price*, but not real output, changes when the *demand curves* shifts; a vertical supply curve that implies fully flexible *prices*.

long-run vertical Phillips Curve The *Phillips Curve* after all *nominal wages* have adjusted to changes in the rate of *inflation*; a line emanating straight upward at the economy's *natural rate of unemployment*.

Lorenz curve A curve showing the distribution of income in an economy. The cumulated percentage of families (income receivers) is measured along the horizontal axis and the cumulated percentage of income is measured along the vertical axis.

loss aversion In *prospect theory*, the property of most people's preferences that the pain generated by losses feels substantially more intense than the pleasure generated by gains.

lump-sum tax A tax that collects a constant amount (the tax revenue of government is the same) at all levels of *GDP*.

***M*1** The most narrowly defined *money supply*, equal to *currency* in the hands of the public and the *checkable deposits* of commercial banks and *thrift institutions*.

***M*2** A more broadly defined *money supply*, equal to *M*1 plus *noncheckable savings accounts* (including *money market deposit accounts*), small *time deposits* (deposits of less than $100,000), and individual *money market mutual fund* balances.

macroeconomics The part of *economics* concerned with the performance and behavior of the economy as a whole. Focuses on *economic growth*, the *business cycle*, *interest rates*, *inflation*, and the behavior of major economic *aggregates* such as the household, business, and government sectors.

managed floating exchange rate An *exchange rate* that is allowed to change (float) as a result of changes in *currency* supply and demand but at times is altered (managed) by governments via their buying and selling of particular currencies.

managerial prerogatives The decisions that a *firm*'s management has the sole right to make; often enumerated in the labor contract (work agreement) between a *labor union* and a firm.

marginal analysis The comparison of *marginal* ("extra" or "additional") *benefits* and *marginal costs*, usually for decision making.

marginal benefit (MB) The extra (additional) benefit of consuming 1 more unit of some *good* or *service;* the change in total benefit when 1 more unit is consumed.

marginal cost (MC) The extra (additional) cost of producing 1 more unit of output; equal to the change in *total cost* divided by the change in output (and, in the short run, to the change in total *variable cost* divided by the change in output).

marginal cost–marginal benefit rule As it applies to *cost-benefit analysis*, the tenet that a government project or program should be expanded to the point where the *marginal cost* and *marginal benefit* of additional expenditures are equal.

marginal product (MP) The additional output produced when 1 additional unit of a resource is employed (the quantity of all other resources employed remaining constant); equal to the change in *total product* divided by the change in the quantity of a resource employed.

marginal productivity theory of income distribution The contention that the distribution of *income* is equitable when each unit of each *resource* receives a *money* payment equal to its *marginal revenue product* (its marginal contribution to the revenue of the *firm* using the unit).

marginal propensity to consume (MPC) The fraction of any change in *disposable income* spent for *consumer goods;* equal to the change in consumption divided by the change in disposable income.

marginal propensity to save (MPS) The fraction of any change in *disposable income* that *households* save; equal to the change in *saving* divided by the change in disposable income.

marginal rate of substitution (MRS) The rate at which a consumer is willing to substitute one good for another (from a given combination of goods) and remain equally satisfied (have the same *total utility*)*;* equal to the slope of a consumer's *indifference curve* at each point on the curve.

marginal resource cost (MRC) The amount by which the total cost of employing a *resource* increases when a *firm* employs

1 additional unit of the resource (the quantity of all other resources employed remaining constant); equal to the change in the *total cost* of the resource divided by the change in the quantity of the resource employed.

marginal revenue The change in *total revenue* that results from the sale of 1 additional unit of a *firm*'s product; equal to the change in total revenue divided by the change in the quantity of the product sold.

marginal-revenue–marginal-cost approach A method that compares the *marginal revenue* and *marginal cost* of each additional unit of output in order to determine the output level that maximizes a firm's *profit* (or minimizes the firm's loss if making a profit is not possible).

marginal revenue product (MRP) The change in a firm's *total revenue* when it employs 1 additional unit of a *resource* (the quantity of all other resources employed remaining constant); equal to the change in *total revenue* divided by the change in the quantity of the resource employed.

marginal revenue productivity (See *marginal revenue product*.)

marginal tax rate The *tax* rate paid on an additional dollar of *income*.

marginal utility The extra *utility* a consumer obtains from the consumption of 1 additional unit of a *good* or *service;* equal to the change in *total utility* divided by the change in the quantity consumed.

market Any institution or mechanism that brings together buyers (demanders) and sellers (suppliers) of a particular *good* or *service*.

market demand (See *total demand*.)

market economy An economy in which *firms* determine how *resources* are allocated; an economy that uses a *market system*.

market failure The inability of a *market* to bring about the allocation of *resources* that best satisfies the wants of society; in particular, the overallocation or underallocation of resources to the production of a particular *good* or *service* because of *externalities* or informational problems or because markets do not provide desired *public goods*.

market for externality rights A *market* in which *firms* can buy and sell pollution permits (which must be obtained by any firm that wishes to discharge pollutants into the environment). The *equilibrium price* determined in the market imposes a cost that discourages pollution and the imposition of *negative externalities* on third parties.

marketing loan program A federal farm subsidy under which certain farmers can receive a loan (on a per-unit-of-output basis) to plant a crop and then, depending on the harvest *price* of the crop, either pay back the loan with interest or keep the loan proceeds while forfeiting their harvested crop to the lender.

market portfolio The portfolio consisting of every financial asset (including every *stock* and *bond*) traded in the financial markets. Used to calculate *beta* (a measure of the degree of riskiness) for specific stocks, bonds, and mutual funds.

market structure The characteristics of an *industry* that define the likely behavior and performance of its *firms*. The primary characteristics are the number of firms in the industry, whether they are selling a *differentiated product*, the ease of entry, and how much control firms have over output prices. The most commonly discussed market structures are *pure competition*, *monopolistic competition*, *oligopoly*, pure *monopoly*, and *monopsony*.

market system (1) An *economic system* in which individuals own most *economic resources* and in which *markets* and *prices* serve as the dominant coordinating mechanism used to allocate those resources; *capitalism*. Compare with *command system*. (2) All the product and resource markets of a *market economy* and the relationships among them.

median-voter model The theory that under majority rule the median (middle) voter will be in the dominant position to determine the outcome of an election.

Medicaid A federal program that helps finance the medical expenses of individuals covered by the *Supplemental Security Income* (*SSI*) and *Temporary Assistance for Needy Families* (*TANF*) programs.

Medicare A federal program that provides for (1) compulsory hospital insurance for senior citizens, (2) low-cost voluntary insurance to help older Americans pay physicians' fees, and (3) subsidized insurance to buy prescription drugs. Financed by *payroll taxes*.

Medicare Part D The portion of Medicare that enables enrollees to shop among private health insurance companies to buy highly subsidized insurance to help reduce the out-of-pocket expense of prescription drugs.

medium of exchange Any item sellers generally accept and buyers generally use to pay for a *good* or *service; money;* a convenient means of exchanging goods and *services* without engaging in *barter*.

mental accounting The tendency people have to create separate "mental boxes" (or "accounts") in which they deal with particular financial transactions in isolation rather than dealing with them as part of an overall decision-making process that would consider how to best allocate their limited budgets across all possible options by using the *utility-maximizing rule*.

menu costs The reluctance of *firms* to cut *prices* during *recessions* (that they think will be short-lived) because of the costs of altering and communicating their price reductions; named after the cost associated with printing new menus at restaurants.

merger The combination of two (or more) *firms* into a single firm.

microeconomics The part of economics concerned with (1) decision making by individual units such as a *household*, a *firm*, or an *industry* and (2) individual markets, specific *goods* and *services*, and product and resource *prices*.

microfinance (Web chapter) The provision of small loans and other financial services to low-income *entrepreneurs* and small-business owners in *developing countries*.

Microsoft case A 2002 antitrust case in which Microsoft was found guilty of violating the *Sherman Act* by engaging in a series of unlawful activities designed to maintain its *monopoly* in operating systems for personal computers; as a remedy the company was prohibited from engaging in a set of specific anticompetitive business practices.

midpoint formula A method for calculating *price elasticity of demand* or *price elasticity of supply* that averages the starting and ending *prices* and quantities when computing percentages.

minimum efficient scale (MES) The lowest level of output at which a *firm* can minimize long-run *average total cost*.

minimum wage The lowest *wage* that employers may legally pay for an hour of work.

mixed economy (See *market system*.)

modern economic growth The historically recent phenomenon in which nations for the first time have experienced sustained increases in *real GDP per capita*.

monetarism The macroeconomic view that the main cause of changes in aggregate output and the *price level* is fluctuations in the *money supply;* espoused by advocates of a *monetary rule*.

monetary multiplier The multiple of its *excess reserves* by which the banking system can expand *checkable deposits* and thus the *money supply* by making new loans (or buying *securities*); equal to 1 divided by the *reserve requirement*.

monetary policy A central bank's changing of the *money supply* to influence *interest* rates and assist the economy in achieving *price-level stability*, *full employment*, and *economic growth*.

monetary rule (1) A set of guidelines to be followed by a *central bank* that wishes to adjust monetary policy over time to achieve goals such as promoting *economic growth*, encouraging *full employment*, and maintaining a stable *price level*. (2) The guidelines for conducting monetary policy suggested by *monetarism*. As traditionally formulated, the *money supply* should be expanded each year at the same annual rate as the potential rate of growth of *real gross domestic product;* the supply of money should be increased steadily between 3 and 5 percent per year. (Also see *Taylor rule*.)

money Any item that is generally acceptable to sellers in exchange for *goods* and *services*.

money capital *Money* available to purchase *capital;* simply *money*, as defined by economists.

money income (See *nominal income*.)

money market The *market* in which the *demand* for and the supply of *money* determine the *interest rate* (or the level of interest rates) in the economy.

money market deposit accounts (MMDAs) Interest-bearing accounts offered by *commercial banks* and *thrift institutions* that invest deposited funds into a variety of short-term *securities*. Depositors may write checks against their balances, but there are minimum-balance requirements as well as limits on the frequency of check writing and withdrawls.

money market mutual funds (MMMFs) *Mutual funds* that invest in short-term *securities*. Depositors can write checks in minimum amounts or more against their accounts.

money supply A nation's supply of *currency* plus *assets* with very high levels of *liquidity*. Different definitions of the money supply include different categories of highly liquid assets. *M*1 uses a more restrictive definition; *M*2, a more broad definition.

monopolistic competition A *market structure* in which many *firms* sell a *differentiated product*, entry is relatively easy, each firm has some control over its product *price*, and there is considerable *nonprice competition*.

monopoly A *market structure* in which there is only a single seller of a good, *service*, or *resource*. In antitrust law, a dominant *firm* that accounts for a very high percentage of total sales within a particular market.

monopsony A *market structure* in which there is only a single buyer of a good, *service*, or *resource*.

moral hazard problem The possibility that individuals or institutions will change their behavior as the result of a contract or agreement. Example: A bank whose deposits are insured against losses may make riskier loans and investments.

mortgage-backed securities *Bonds* that represent claims to all or part of the monthly mortgage payments from the pools of mortgage loans made by leaders to borrowers to help them purchase residential property.

mortgage debt crisis The period beginning in late 2007 when thousands of homeowners defaulted on mortgage loans when they experienced a combination of higher mortgage interest rates and falling home *prices*.

MR = MC rule The principle that a *firm* will maximize its profit (or minimize its losses) by producing the output at which *marginal revenue* and *marginal cost* are equal, provided product *price* is equal to or greater than *average variable cost*.

MRP = MRC rule The principle that to maximize profit (or minimize losses), a *firm* should employ the quantity of a resource at which its *marginal revenue product* (MRP) is equal to its *marginal resource cost* (MRC), the latter being the wage rate in a purely competitive labor market.

multinational corporations Firms that own production facilities in two or more countries and produce and sell their products globally.

multiple counting Wrongly including the value of *intermediate goods* in the *gross domestic product;* counting the same *good* or *service* more than once.

multiplier The ratio of a change in *equilibrium GDP* to the change in *investment* or in any other component of *aggregate expenditures* or *aggregate demand;* the number by which a change in any such component must be multiplied to find the resulting change in equilibrium GDP.

multiplier effect The effect on *equilibrium GDP* of a change in *aggregate expenditures* or *aggregate demand* (caused by a change in the *consumption schedule*, *investment*, *government purchases*, or *net exports*).

mutual funds *Portfolios* of *stocks* and *bonds* selected and purchased by mutual fund companies, which finance the purchases by pooling money from thousands of individual investors; includes both *index funds* as well as *actively managed funds*. Fund returns (profits or losses) pass through to each fund's investors.

mutual interdependence A situation in which a change in *price* strategy (or in some other strategy) by one *firm* will affect the sales and profits of another firm (or other firms). Any firm that makes such a change can expect its rivals to react to the change.

myopia Refers to the difficulty human beings have with conceptualizing the more distant future. Leads to decisions that overly favor present and near-term options at the expense of more distant future possibilities.

Nash equilibrium In *game theory*, an outcome from which neither of two rival *firms* wishes to deviate; the outcome, once achieved, is stable and therefore lasting.

national bank In the United States, a *commercial bank* authorized to operate by the federal government.

National Credit Union Administration (NCUA) The federally chartered agency that insures *credit union* deposit liabilities (up to $250,000 per account).

national health insurance A program in which a nation's government provides a basic package of health care to all citizens at no direct charge or at a low cost-sharing level. Financing is out of general *tax* revenues.

national income Total *income* earned by *resource* suppliers for their contributions to *gross domestic product* plus *taxes on production and imports;* the sum of wages and salaries, *rent*, *interest*, *profit*, *proprietors' income*, and such taxes.

national income accounting The techniques used to measure the overall production of a country's economy as well as other related variables.

National Labor Relations Act (NLRA) The basic labor-relations law in the United States. Defines the legal rights of unions and management and identifies unfair union and management labor practices; established the *National Labor Relations Board*. Often referred to as the Wagner Act, after the legislation's sponsor, New York Senator Robert F. Wagner.

National Labor Relations Board (NLRB) The board established by the *National Labor Relations Act* of 1935 to investigate unfair labor practices, issue *cease-and-desist orders*, and conduct elections among employees to determine if they wish to be represented by a *labor union*.

natural monopoly An *industry* in which *economies of scale* are so great that a single *firm* can produce the industry's product at a lower average total cost than would be possible if more than one firm produced the product.

natural rate of unemployment (NRU) The *full-employment rate of unemployment;* the *unemployment rate* occurring when there is no cyclical unemployment and the economy is achieving its *potential output*; the unemployment rate at which actual *inflation* equals expected inflation.

near-money Financial *assets* that are not themselves a *medium of exchange* but that have extremely high *liquidity* and thus can be readily converted into *money*. Includes noncheckable *savings accounts*, *time deposits*, and short-term *U.S. government securities* plus savings bonds.

negative externality A cost imposed without compensation on third parties by the production or consumption of sellers or buyers. Example: A manufacturer dumps toxic chemicals into a river, killing fish prized by sports fishers; an external cost or a spillover cost.

negative GDP gap A situation in which actual *gross domestic product* is less than *potential output*. Also known as a recessionary output gap.

negative relationship (See *inverse relationship*.)

negative self-selection As it relates to international migration, the idea that those who choose to move to another country have poorer *wage* opportunities in the origin country than those with similar skills who choose not to *emigrate*.

negative-sum game In *game theory*, a game in which the gains (+) and losses (−) add up to some amount less than zero; one party's losses exceed the other party's gains.

neoclassical economics The dominant and conventional branch of economic theory that attempts to predict human behavior by building economic models based on simplifying assumptions about people's motives and capabilities. These include that people are fundamentally *rational;* motivated almost entirely by *self-interest;* good at math; and unaffected by *heuristics*, *time inconsistency*, and *self-control problems*.

net benefits The total benefits of some activity or policy less the total costs of that activity or policy.

net domestic product (NDP) *Gross domestic product* less the part of the year's output that is needed to replace the *capital goods* worn out in producing the output; the nation's total output available for consumption or additions to the *capital stock*.

net exports (X_n) *Exports* minus *imports*.

net foreign factor income Receipts of *resource* income from the rest of the world minus payments of resource income to the rest of the world.

net investment income The *interest* and *dividend* income received by the residents of a nation from residents of other nations less the interest and dividend payments made by the residents of that nation to the residents of other nations.

net private domestic investment *Gross private domestic investment* less *consumption of fixed capital;* the addition to the nation's stock of *capital* during a year.

net taxes The *taxes* collected by government less *government transfer payments*.

net transfers The personal and government *transfer payments* made by one nation to residents of foreign nations less the personal and government transfer payments received from residents of foreign nations.

network effects Increases in the value of a product to each user, including existing users, as the total number of users rises.

net worth The total *assets* less the total *liabilities* of a *firm* or an individual; for a firm, the claims of the owners against the firm's total assets; for an individual, his or her wealth.

new classical economics The theory that, although unanticipated *price-level* changes may create macroeconomic instability in the short run, the economy will return to and stabilize at the full-employment level of domestic output in the long run because *prices* and *wages* adjust automatically to correct movements away from the full-employment output level.

NLRB (See *National Labor Relations Board*.)

nominal gross domestic product (GDP) *GDP* measured in terms of the *price level* at the time of measurement; *GDP* not adjusted for *inflation*.

nominal income The number of dollars received by an individual or group for its *resources* during some period of time.

nominal interest rate The *interest rate* expressed in terms of annual amounts currently charged for *interest* and not adjusted for *inflation*.

nominal wage The amount of *money* received by a worker per unit of time (hour, day, etc.); money wage.

noncash transfer A *government transfer payment* in the form of *goods* and *services* rather than *money*, for example, food stamps, housing assistance, and job training; also called *in-kind transfers*.

noncollusive oligopoly An *oligopoly* in which the *firms* do not act together and in agreement to determine the *price* of the *industry*'s product or the output that each firm will produce.

noncompeting groups Collections of workers who do not compete with each other for employment because the skill and training of the workers in one group are substantially different from those of the workers in other groups.

nondiscretionary fiscal policy (See *built-in stabilizer*.)

nondiversifiable risk Investment *risk* that investors are unable to reduce via *diversification;* also called systemic risk.

nondurable good A *consumer good* with an expected life (use) of less than three years.

nonexcludability The inability to keep nonpayers (free riders) from obtaining benefits from a certain good; a characteristic of a *public good*.

nonexhaustive expenditure An expenditure by government that does not result directly in the use of economic resources or the production of *goods* and *services*; see *government transfer payment*.

nonincome determinants of consumption and saving All influences on consumption and *saving* other than the level of *GDP*.

noninterest determinants of investment All influences on the level of *investment* spending other than the *interest rate*.

noninvestment transaction An expenditure to purchase financial *assets* such as *stocks* and *bonds* or to purchase secondhand *capital goods*; any *financial investment*. By contrast, *investment* is spending for the production of new *capital goods*.

nonmarket transactions The value of the *goods* and *services* that are not included in the *gross domestic product* because they are not bought and sold.

nonprice competition Competition based on distinguishing one's product by means of *product differentiation* and then *advertising* the distinguished product to consumers.

nonproduction transaction The purchase and sale of any item that is not a currently produced *good* or *service*.

nonrenewable natural resource Things such as oil, natural gas, and metals, that are either in actual fixed supply or that renew so slowly as to be in virtual fixed supply when viewed from a human time perspective.

nonrivalry The idea that one person's benefit from a certain *good* does not reduce the benefit available to others; a characteristic of a *public good*.

nontariff barriers (NTBs) All barriers other than *protective tariffs* that nations erect to impede international trade, including *import quotas*, licensing requirements, unreasonable product-quality standards, unnecessary bureaucratic detail in customs procedures, and so on.

normal good A *good* or *service* whose consumption increases when *income* increases and falls when income decreases, *price* remaining constant.

normal profit The payment made by a *firm* to obtain and retain *entrepreneurial ability*; the minimum *income* that entrepreneurial ability must receive to induce *entrepreneurs* to provide their entrepreneurial ability to a firm; the level of *accounting profit* at which a firm generates an *economic profit* of zero after paying for entrepreneurial ability.

normative economics The part of economics involving value judgments about what the economy should be like; focused on which economic goals and policies should be implemented; policy economics.

North American Free Trade Agreement (NAFTA) The 1993 treaty that established an international free-trade zone composed of Canada, Mexico, and the United States.

occupation A category of paid labor employment defined by its activities or tasks rather than by the particular employer or *industry*. Examples are managers, nurses, farmers, and cooks.

occupational licensing The laws of state or local governments that require that a worker satisfy certain specified requirements and obtain a license from a licensing board before engaging in a particular occupation.

occupational segregation The crowding of women or minorities into less desirable, lower-paying occupations.

official reserves Foreign *currencies* owned by the central bank of a nation.

offshoring The practice of shifting work previously done by domestic workers to workers located abroad.

Okun's law The generalization that any 1-percentage-point rise in the *unemployment rate* above the *full-employment rate of unemployment* is associated with a rise in the *negative GDP gap* by 2 percent of *potential output* (potential *GDP*).

oligopoly A *market structure* in which a few *firms* sell either a *standardized* or *differentiated product*, into which entry is difficult, in which the firm has limited control over product *price* because of *mutual interdependence* (except when there is collusion among firms), and in which there is typically *nonprice competition*.

one-time game In *game theory*, a game in which the parties select their optimal strategies in a single time period without regard to possible interaction in subsequent time periods.

OPEC (See *Organization of Petroleum Exporting Countries*.)

open economy An economy that exports and imports *goods* and *services*.

open-market operations The purchases and sales of U.S. government *securities* that the *Federal Reserve System* undertakes in order to influence *interest rates* and the *money supply;* one method by which the *Federal Reserve* implements *monetary policy*.

open shop A place of employment in which the employer may hire nonunion workers and in which the workers need not become members of a *labor union*.

opportunity cost The amount of other products that must be forgone or sacrificed to produce a unit of a product.

opportunity-cost ratio An equivalency showing the number of units of two products that can be produced with the same *resources;* the equivalency 1 corn ≡ 3 olives shows that the resources required to produce 3 units of olives must be shifted to corn production to produce 1 unit of corn.

optimal amount of R&D (Web chapter) The level of *R&D* at which the *marginal benefit* and *marginal cost* of R&D expenditures are equal.

optimal reduction of an externality The reduction of a *negative externality* such as pollution to the level at which the *marginal benefit* and *marginal cost* of reduction are equal.

ordinal utility Satisfaction that is measured by having consumers compare and rank products (or combinations of products) as to preference, without asking them to specify the absolute amounts of satisfaction provided by the products.

Organization of Petroleum Exporting Countries (OPEC) A cartel of 12 oil-producing countries (Algeria, Angola, Ecuador, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, Venezuela, and the United Arab Emirates) that attempts to control the quantity and *price* of crude oil exported by its members and that accounts for a large percentage of the world's export of oil.

other-things-equal assumption The assumption that factors other than those being considered are held constant; *ceteris paribus* assumption.

outpayments The expenditures of domestic or foreign *currency* that the individuals, *firms*, and governments of one nation make to purchase *goods* and *services*, for remittances, to pay investment *income*, and for purchases of foreign *assets*.

output effect The possibility that when the *price* of the first of a pair of *substitute resources* falls, the *quantity demanded* of both resources will rise because the reduction in the price of the first resource so greatly reduces production costs that the volume of output created with the two resources increases by so much that the quantity demanded of the second resource increases even after accounting for the *substitution effect*. (See the second definition listed in the entry for *substitution effect*.)

***P* = MC rule** The principle that a purely competitive *firm* will maximize its profit or minimize its loss by producing that output at which the *price* of the product is equal to *marginal cost*, provided that price is equal to or greater than *average variable cost* in the short run and equal to or greater than *average total cost* in the long run.

paper money Pieces of paper used as a *medium of exchange;* in the United States, *Federal Reserve Notes*.

paradox of thrift The seemingly self-contradictory but possibly true statement that increased *saving* may be both good and bad for the economy. It is always good in the long run when matched with increased *investment* spending, but may be bad in the short run if there is a *recession* because it reduces spending on *goods* and *services*. If the increased savings are not translated into increased investment, then the fall in consumption spending will not be made up for by an increase in investment. The overall result will be a decrease in output and employment. If the decline in *GDP* is severe enough, the attempt to save more will actually lead to less overall savings because the higher rate of saving will be applied to a smaller *national income*. Attempts by *households* to save more during a recession may simply worsen the recession and result in less saving.

paradox of voting A situation where paired-choice voting by majority rule fails to provide a consistent ranking of society's preferences for *public goods* or *public services*.

parity concept The idea that year after year the sale of a specific output of a farm product should enable a farmer to purchase a constant amount of nonagricultural *goods* and *services*.

parity ratio The ratio of the *price* received by farmers from the sale of an agricultural commodity to the prices of other goods paid by them; usually expressed as a percentage; used as a rationale for *price supports*.

partnership An unincorporated *firm* owned and operated by two or more persons.

passively managed funds *Mutual funds* whose *portfolios* are not regularly updated by a fund manager attempting to generate high returns. Rather, once an initial portfolio is selected, it is left unchanged so that investors receive whatever return that unchanging portfolio subsequently generates. *Index funds* are a type of passively managed fund.

patent (Web chapter) An exclusive right given to inventors to produce and sell a new product or machine for 20 years from the time of patent application.

Patient Protection and Affordable Care Act (PPACA) A major health care law passed by the federal government in 2010. Major provisions include an individual health insurance mandate, a ban on insurers refusing to accept patients with preexisting conditions, and federal (rather than state) regulation of health insurance policies.

payments deficit (See *balance-of-payments deficit*.)

payments surplus (See *balance-of-payments surplus*.)

payroll tax A *tax* levied on employers of labor equal to a percentage of all or part of the *wages* and salaries paid by them and on employees equal to a percentage of all or part of the wages and salaries received by them.

peak The point in a *business cycle* at which business activity has reached a temporary maximum; the point at which an *expansion* ends and a *recession* begins. At the peak, the economy is near or at *full employment* and the level of real output is at or very close to the economy's capacity.

per capita GDP *Gross domestic product* (*GDP*) per person; the average GDP of a population.

per capita income A nation's total *income* per person; the average income of a population.

percentage rate of return The percentage gain or loss, relative to the buying *price*, of an *economic investment* or *financial investment* over some period of time.

perfectly elastic demand Product or *resource* demand in which *quantity demanded* can be of any amount at a particular product or resource *price;* graphs as a horizontal *demand curve*.

perfectly elastic supply Product or *resource* supply in which *quantity supplied* can be of any amount at a particular product or resource *price;* graphs as a horizontal *supply curve*.

perfectly inelastic demand Product or *resource* demand in which *price* can be of any amount at a particular quantity of the product or resource that is demanded; when the *quantity demanded* does not respond to a change in price; graphs as a vertical *demand curve*.

perfectly inelastic supply Product or resource supply in which *price* can be of any amount at a particular quantity of the product or resource that is demanded; when the *quantity supplied* does not respond to a change in price; graphs as a vertical *supply curve*.

per se violations Collusive actions, such as attempts by *firms* to fix *prices* or divide a market, that are violations of the *antitrust laws,* even if the actions themselves are unsuccessful.

personal consumption expenditures (*C*) The expenditures of *households* for both durable and nondurable *consumer goods*.

personal distribution of income The manner in which the economy's *personal* or *disposable income* is divided among different *income* classes or different *households* or families.

personal income (PI) The earned and unearned *income* available to resource suppliers and others before the payment of personal *taxes*.

personal income tax A *tax* levied on the taxable income of individuals, *households*, and unincorporated *firms*.

personal mandate The requirement under the *Patient Protection and Affordable Care Act* (*PPACA*) of 2010 that all U.S. citizens and legal residents purchase health insurance unless they are already covered by employer-sponsored health insurance or government-sponsored health insurance (*Medicaid* or *Medicare*).

personal saving The *personal income* of *households* less personal *taxes* and *personal consumption expenditures; disposable income* not spent for *consumer goods*.

per-unit production cost The average production cost of a particular level of output; total input cost divided by units of output.

Phillips Curve A curve showing the relationship between the *unemployment rate* (on the horizontal axis) and the annual rate of increase in the *price level* (on the vertical axis).

planned investment The amount that *firms* plan or intend to invest.

plant A physical establishment that performs one or more functions in the production, fabrication, and distribution of *goods* and *services*.

policy economics The formulation of courses of action to bring about desired economic outcomes or to prevent undesired occurrences.

political business cycle Fluctuations in the economy caused by the alleged tendency of Congress to destabilize the economy by reducing taxes and increasing government expenditures before elections and to raise taxes and lower expenditures after elections.

political corruption The unlawful misdirection of governmental resources or actions that occurs when government officials abuse their entrusted powers for personal gain. (Also see *corruption*.)

Ponzi scheme A financial fraud in which the returns paid to earlier investors come from contributions made by later investors (rather than from the *financial investment* that the perpetrator of the fraud claims to be making). Named after notorious fraudster Charles Ponzi.

portfolio A specific collection of *stocks*, *bonds*, or other *financial investments* held by an individual or a *mutual fund*.

positive economics The analysis of facts or data to establish scientific generalizations about economic behavior.

positive externality A benefit obtained without compensation by third parties from the production or consumption of sellers or buyers. Example: A beekeeper benefits when a neighboring farmer plants clover. An *external benefit* or a spillover benefit.

positive GDP gap A situation in which actual *gross domestic product* exceeds *potential output*. Also known as an *inflationary expenditure* gap.

positive relationship (See *direct relationship*.)

positive sum game In *game theory*, a game in which the gains (+) and losses (−) add up to more than zero; one party's gains exceed the other party's losses.

***post hoc, ergo propter hoc* fallacy** The false belief that when one event precedes another, the first event must have caused the second event.

potential competition The new competitors that may be induced to enter an *industry* if *firms* now in that industry are receiving large *economic profits*.

potential output The real output (*GDP*) an economy can produce when it fully employs its available resources.

poverty A situation in which the basic needs of an individual or family exceed the means to satisfy them.

poverty rate The percentage of the population with incomes below the official poverty income levels that are established by the federal government.

precommittments Actions taken ahead of time that make it difficult for the future self to avoid doing what the present self desires. See *time inconsistency* and *self-control problems*.

preferred provider organization (PPO) An arrangement in which doctors and hospitals agree to provide health care to insured individuals at rates negotiated with an insurer.

present value Today's value of some amount of *money* that is to be received sometime in the future.

price The amount of *money* needed to buy a particular *good*, *service*, or *resource*.

price ceiling A legally established maximum *price* for a *good*, or *service*. Normally set at a price below the *equilibrium price*.

price discrimination The selling of a product to different buyers at different *prices* when the price differences are not justified by differences in cost.

price elasticity of demand The ratio of the percentage change in *quantity demanded* of a product or *resource* to the percentage change in its *price;* a measure of the responsiveness of buyers to a change in the price of a product or resource.

price elasticity of supply The ratio of the percentage change in *quantity supplied* of a product or *resource* to the percentage change in its *price;* a measure of the responsiveness of producers to a change in the price of a product or resource.

price fixing The conspiring by two or more *firms* to set the *price* of their products; an illegal practice under the *Sherman Act*.

price floor A legally established minimum *price* for a *good*, or service. Normally set at a price above the *equilibrium price*.

price index An index number that shows how the weighted-average *price* of a "market basket" of goods changes over time relative to its price in a specific *base year*.

price leadership An informal method that *firms* in an *oligopoly* may employ to set the *price* of their product: One firm (the leader) is the first to announce a change in price, and the other firms (the followers) soon announce identical or similar changes.

price level The weighted average of the *prices* of all the *final goods* and final *services* produced in an economy.

price-level stability A steadiness of the *price level* from one period to the next; zero or low annual *inflation;* also called "price stability."

price-level surprises Unanticipated changes in the *price level*.

price maker A seller (or buyer) that is able to affect the product or resource *price* by changing the amount it sells (or buys).

price support The term used to refer to *price floors* applied to *farm commodities;* the minimum *price* that the government allows farmers to receive for farm commodities like wheat or corn.

price taker A seller (or buyer) that is unable to affect the *price* at which a product or *resource* sells by changing the amount it sells (or buys).

price war Successive, competitive, and continued decreases in the *prices* charged by *firms* in an oligopolistic *industry*. At each stage of the price war, one *firm* lowers its price below its rivals' price, hoping to increase its sales and revenues at its rivals' expense. The war ends when the price decreases cease.

prime interest rate The benchmark *interest rate* that banks use as a reference point for a wide range of loans to businesses and individuals.

principal-agent problem (1) At a *firm*, a conflict of interest that occurs when agents (workers or managers) pursue their own objectives to the detriment of the principals' (stockholders') goals. (2) In *public choice theory*, a conflict of interest that arises when elected officials (who are the agents of the people) pursue policies that are in their own interests rather than policies that would be in the better interests of the public (the principals).

principle of comparative advantage The proposition that an individual, region, or nation will benefit if it specializes in producing goods for which its own *opportunity costs* are lower than the opportunity costs of a trading partner, and then exchanging some of the products in which it specializes for other desired products produced by others.

private good A *good*, or *service* that is individually consumed and that can be profitably provided by privately owned *firms* because they can exclude nonpayers from receiving the benefits.

private property The right of private persons and *firms* to obtain, own, control, employ, dispose of, and bequeath *land*, *capital*, and other property.

private sector The *households* and business *firms* of the economy.

probability-weighted average Each of the possible future rates of return from an investment multiplied by its respective probability (expressed as a decimal) of happening.

process innovation (Web chapter) The development and use of new or improved production or distribution methods.

producer surplus The difference between the actual *price* a producer receives (or producers receive) and the minimum acceptable price; the triangular area above the *supply curve* and below the market price.

product differentiation A strategy in which one *firm*'s product is distinguished from competing products by means of its design, related *services*, quality, location, or other attributes (except *price*).

product innovation (Web chapter) The development and sale of a new or improved product (or service).

production possibilities curve A curve showing the different combinations of two goods or *services* that can be produced in a *full-employment*, *full-production* economy where the available supplies of *resources* and technology are fixed.

productive efficiency The production of a *good* in the least costly way; occurs when production takes place at the output at which *average total cost* is a minimum and *marginal product* per dollar's worth of input is the same for all inputs.

productivity A measure of average output or real output per unit of input. For example, the productivity of labor is determined by dividing real output by hours of work.

productivity growth The increase in *productivity* from one period to another.

product market A market in which products are sold by *firms* and bought by *households*.

profit Usually refers to *economic profit* (*total revenue* minus both *explicit costs* and *implicit costs*) but may also refer to *accounting profit* (total revenue minus just explicit costs). If it refers to economic profit, then it is the return earned by the *resource* known as *entrepreneurial ability*.

profit-maximizing combination of resources The quantity of each *resource* a *firm* must employ to maximize its *profit* or minimize its loss; the combination of resource inputs at which the *marginal revenue product* of each resource is equal to its *marginal resource cost* (to its *price* if the resource is employed in a competitive market).

profit-sharing plan A compensation device through which workers receive part of their pay in the form of a share of their employer's *profit* (if any).

progressive tax At the individual level, a *tax* whose *average tax rate* increases as the taxpayer's *income* increases. At the national level, a *tax* for which the *average tax rate* (= tax revenue/*GDP*) rises with *GDP*.

property tax A *tax* on the value of property (*capital*, *land*, *stocks* and *bonds*, and other *assets*) owned by *firms* and *households*.

proportional tax At the individual level, a *tax* whose *average tax rate* remains constant as the taxpayer's *income* increases or decreases. At the national level, a *tax* for which the *average tax rate* (= tax revenue/*GDP*) remains constant as *GDP* rises or falls.

proprietor's income The net income (*profit*) of the owners of unincorporated *firms* (*proprietorships* and *partnerships*).

prospect theory A *behavioral economics* theory of preferences having three main features: (1) people evaluate options on the basis of whether they generate gains or losses relative to the *status quo;* (2) gains are subject to *diminishing marginal utility*, while losses are subject to diminishing marginal disutility; and (3) people are prone to *loss aversion*.

protective tariff A *tariff* designed to shield domestic producers of a *good* or *service* from the competition of foreign producers.

public assistance programs Government programs that pay benefits to those who are unable to earn *income* (because of permanent disabilities or because they have very low income and dependent children); financed by general *tax* revenues and viewed as public charity (rather than earned rights).

public choice theory The economic analysis of government decision making, politics, and elections.

public debt The total amount owed by the federal government to the owners of government *securities;* equal to the sum of past government *budget deficits* less government *budget surpluses*.

public good A *good* or *service* that is characterized by *nonrivalry* and *nonexcludability*. These characteristics typically imply that no private *firm* can break even when attempting to provide such products. As a result, they are often provided by governments, who pay for them using general *tax* revenues.

public interest theory of regulation The presumption that the purpose of the regulation of an *industry* is to protect the public (consumers) from abuse of the power possessed by *natural monopolies*.

public investments Government expenditures on public capital (such as roads, highways, bridges, mass-transit systems, and electric power facilities) and on *human capital* (such as education, training, and health).

public sector The part of the economy that contains all government entities; government.

public utility A *firm* that produces an essential *good* or *service*, has obtained from a government the right to be the sole supplier of the good or service in an area, and is regulated by that government to prevent the abuse of its *monopoly* power.

purchasing power The amount of *goods* and *services* that a monetary unit of *income* can buy.

purchasing power parity The idea that if countries have *flexible exchange rates* (rather than *fixed exchange rates*), the exchange rates between national currencies will adjust to equate the purchasing power of various currencies. In particular, the exchange rate between any two national currencies will adjust to reflect the *price-level* differences between the two countries.

pure competition A *market structure* in which a very large number of *firms* sells a *standardized product*, into which entry is very easy, in which the individual seller has no control over the product *price*, and in which there is no nonprice competition; a market characterized by a very large number of buyers and sellers.

purely competitive labor market A *resource market* in which many *firms* compete with one another in hiring a specific kind of *labor*, numerous equally qualified workers supply that labor, and no one controls the market *wage rate*.

pure monopoly A *market structure* in which one *firm* sells a unique product, into which entry is blocked, in which the single firm has considerable control over product *price*, and in which *nonprice competition* may or may not be found.

pure profit (See *economic profit*.)

pure rate of interest The hypothetical *interest rate* that is completely *risk*-free and not subject to market imperfections; the hypothetical interest rate that would only compensate investors for *time preference*.

quantitative easing (QE) An *open-market operation* in which *bonds* are purchased by a *central bank* in order to increase the quantity of *excess reserves* held by *commercial banks* and thereby (hopefully) stimulate the economy by increasing the amount of lending undertaken by commercial banks; undertaken when interest rates are near zero and, consequently, it is not possible for the central bank to further stimulate the economy with lower interest rates due to the *zero lower bound problem*.

quantity demanded The amount of a *good* or *service* that buyers (or a buyer) are willing and able to purchase at a specific *price* during a specified period of time.

quantity supplied The amount of a *good* or *service* that producers (or a producer) are willing and able to make available for sale at a specific *price* during a specified period of time.

quasi-public bank A bank that is privately owned but governmentally (publicly) controlled; each of the U.S. *Federal Reserve Banks*.

quasi-public good A *good* or *service* to which *excludability* could apply but that has such a large *positive externality* that government sponsors its production to prevent an underallocation of resources.

R&D Research and development activities undertaken to bring about *technological advance*.

rate of exchange The *price* paid in one's own *money* to acquire 1 unit of a foreign *currency;* the rate at which the money of one nation is exchanged for the money of another nation.

rate of return The gain in net revenue divided by the cost of an *investment* or an *R&D* expenditure; often expressed as a *percentage rate of return*.

rational Behaviors and decisions that maximize a person's chances of achieving his or her goals. See *rational behavior*.

rational behavior Human behavior based on comparison of *marginal costs* and *marginal benefits;* behavior designed to maximize *total utility*. See *rational*.

rational expectations theory The hypothesis that *firms* and *households* expect monetary and fiscal policies to have certain effects on the economy and (in pursuit of their own self-interests) take actions that make these policies ineffective.

rationing function of prices The ability of *market* forces in competitive markets to equalize *quantity demanded* and *quantity supplied* and to eliminate *shortages* and *surpluses* via changes in *prices*.

real-balances effect The tendency for increases in the *price level* to lower the real value (or *purchasing power*) of financial *assets* with fixed *money* value and, as a result, to reduce total spending and real output, and conversely for decreases in the *price level*.

real-business-cycle theory A theory that *business cycles* result from changes in *technology* and *resource* availability, which affect *productivity* and thus increase or decrease long-run *aggregate supply*.

real capital (See *capital*.)

real GDP (See *real gross domestic product*.)

real GDP per capita *Inflation*-adjusted output per person; *real GDP*/population.

real gross domestic product (GDP) *Gross domestic product* adjusted for *inflation;* gross domestic product in a year divided by the GDP *price index* for that year, the index expressed as a decimal.

real income The amount of *goods* and *services* that can be purchased with *nominal income* during some period of time; nominal income adjusted for *inflation*.

real interest rate The *interest rate* expressed in dollars of constant value (adjusted for *inflation*) and equal to the *nominal interest rate* less the expected rate of inflation.

real wage The amount of *goods* and *services* a worker can purchase with his or her *nominal wage;* the *purchasing power* of the *nominal wage*.

recession A period of declining *real GDP*, accompanied by lower *real income* and higher *unemployment*.

recessionary expenditure gap The amount by which the *aggregate expenditures schedule* must shift upward to increase *real GDP* to its full-employment, noninflationary level.

regressive tax At the individual level, a *tax* whose *average tax rate* decreases as the taxpayer's *income* increases. At the national level, a *tax* for which the *average tax rate* (= tax revenue/*GDP*) falls as *GDP* rises.

regulatory agency An agency, commission, or board established by the federal government or a state government to control the *prices* charged and the *services* offered by a *natural monopoly* or *public utility*.

regulatory capture The situation that occurs when a governmental *regulatory agency* ends up being controlled by the industry that it is supposed to be regulating.

remittances Payments by *immigrants* to family members and others located in the immigrants' home countries.

renewable natural resources Things such as forests, water in reservoirs, and wildlife that are capable of growing back or building back up (renewing themselves) if they are harvested at moderate rates.

rental income The payments received by those who supply *land* to the economy.

rent-seeking behavior The actions by persons, *firms*, or unions to gain special benefits from government at the taxpayers' or someone else's expense.

repeated game In *game theory*, a game that is played again sometime after the previous game ends.

replacement rate The *total fertility rate* necessary to offset deaths in a country and thereby keep the size of its population constant (without relying on immigration). For most countries, a total fertility rate of about 2.1 births per woman per lifetime.

required reserves The funds that each *commercial bank* and *thrift institution* must deposit with its local *Federal Reserve Bank* (or hold as *vault cash*) to meet the legal *reserve requirement;* a fixed percentage of each bank's or thrift's *checkable deposits*.

reserve ratio The fraction of *checkable deposits* that each *commercial bank* or *thrift institution* must hold as reserves at its local *Federal Reserve Bank* or in its own bank vault; also called the *reserve requirement*.

reserve requirement The specified minimum percentage of its *checkable deposits* that each *commercial bank* or *thrift institution* must keep on deposit at the *Federal Reserve Bank* in its district or hold as *vault cash*.

resource A natural, human, or manufactured item that helps produce *goods* and *services;* a productive agent or factor of production; one of the four *economic resources*.

resource market A market in which *households* sell and *firms* buy *resources* or the services of resources.

restrictive monetary policy *Federal Reserve System* actions to reduce the *money supply*, increase *interest rates*, and reduce *inflation;* a tight money policy.

revenue tariff A *tariff* designed to produce *income* for the federal government.

right-to-work law A state law (in 23 states) that makes it illegal to require that a worker join a *labor union* in order to retain his or her job; laws that make *union shops* and *agency shops* illegal.

risk The uncertainty as to the future returns of a particular *financial investment* or *economic investment*.

risk-free interest rate The *interest rate* earned on short-term U.S. government *bonds*.

risk premium The *interest rate* above the *risk-free* interest rate that must be paid and received to compensate a lender or investor for *risk*.

rivalry (1) The characteristic of a *private good*, the consumption of which by one party excludes other parties from obtaining the benefit; (2) the attempt by one *firm* to gain strategic advantage over another firm to enhance market share or *profit*.

rule of reason The rule stated and applied in the *U.S. Steel case* that only combinations and contracts unreasonably restraining trade are subject to actions under the antitrust laws and that size and possession of *monopoly* power are not by themselves illegal. Compare with *per se violation*.

rule of 70 A method for determining the number of years it will take for some measure to double, given its annual percentage increase. Example: To determine the number of years it will take for the *price level* to double, divide 70 by the annual rate of *inflation*.

sales and excise taxes (See *sales tax;* see *excise tax*.)

sales tax A *tax* levied on the cost (at retail) of a broad group of products.

saving *Disposable income* not spent for *consumer goods;* equal to *disposable income* minus *personal consumption expenditures;* saving is a flow. Compare with *savings*.

savings The accumulation of funds that results when people in an economy spend less (consume less) than their *incomes* during a given time period; savings are a stock. Compare with *saving*.

savings account A deposit in a *commercial bank* or *thrift institution* on which *interest* payments are received; generally used for *saving* rather than daily transactions; a component of the *M2 money supply*.

savings and loan association (S&L) A *firm* that accepts deposits primarily from small individual savers and lends primarily to individuals to finance purchases such as autos and homes; now nearly indistinguishable from a *commercial bank*.

saving schedule A table of numbers that shows the amounts *households* plan to save (plan not to spend for *consumer goods*), at different levels of *disposable income*.

savings deposit A deposit that is *interest*-bearing and that the depositor can normally withdraw at any time.

savings institution (See *thrift institution*.)

Say's law The largely discredited macroeconomic generalization that the production (*supply*) of *goods* and *services* creates an equal *demand* for those goods and services.

scarce resources The limited quantities of *land*, *capital*, *labor*, and *entrepreneurial ability* that are never sufficient to satisfy people's virtually unlimited economic wants.

scarcity The limits placed on the amounts and types of *goods* and *services* available for consumption as the result of there being only limited *economic resources* from which to produce output; the fundamental economic constraint that creates *opportunity costs* and that necessitates the use of *marginal analysis* (*cost-benefit analysis*) to make optimal choices.

scientific method The procedure for the systematic pursuit of knowledge involving the observation of facts and the formulation and testing of hypotheses to obtain theories, principles, and laws.

secular trend A long-term tendency; a change in some variable over a very long period of years.

securities *Stocks*, *bonds*, and other financial documents that attest to a financial obligation.

securitization The process of aggregating many individual financial debts, such as mortgages or student loans, into a pool and then issuing new *securities* (typically *bonds*) backed by the pool. The holders of the new securities are entitled to receive the debt payments made on the individual financial debts in the pool.

Security Market Line (SML) A line that shows the *average expected rate of return* of all *financial investments* at each level of *nondiversifiable risk*, the latter measured by *beta*.

self-control problems Refers to the difficulty people have in sticking with earlier plans and avoiding suboptimal decisions when finally confronted with a particular decision-making situation. A manifestation of *time inconsistency* and potentially avoidable by using *precommitments*.

self-interest That which each *firm*, property owner, worker, and consumer believes is best for itself and seeks to obtain.

self-selection As it relates to international migration, the idea that those who choose to move to a new country tend to have greater motivation for economic gain or greater willingness to sacrifice current consumption for future consumption than those with similar skills who choose to remain at home.

seniority The length of time a worker has been employed absolutely or relative to the other workers at the same *firm;* may be used to determine which workers will be laid off (if there is insufficient work to employ all of the workers at the firm) as well as who will be the first to be rehired (if and when more work becomes available again at the firm).

separation of ownership and control The fact that different groups of people own a *corporation* (the stockholders) and manage it (the directors and officers).

sequential game In *game theory*, a game in which the parties make their moves in turn, with one party making the first move, followed by the other party making the next move, and so on.

service An (intangible) act or use for which a consumer, *firm*, or government is willing to pay.

Sherman Act The federal antitrust law of 1890 that makes *monopoly* and conspiracies to restrain trade criminal offenses.

shirking Workers' neglecting or evading *work* to increase their *utility* or well-being.

shocks Sudden, unexpected changes in *demand* (or *aggregate demand*) or supply (or *aggregate supply*).

shortage The amount by which the *quantity demanded* of a product exceeds the *quantity supplied* at a particular (below-equilibrium) *price*.

short run (1) In *microeconomics*, a period of time in which producers are able to change the quantities of some but not all of the *resources* they employ; a period in which some resources (usually *plant*) are fixed and some are variable. (2) In *macroeconomics*, a period in which *nominal wages* and other input *prices* do not change in response to a change in the *price level*.

short-run aggregate supply curve An *aggregate supply* curve relevant to a time period in which input *prices* (particularly *nominal wages*) do not change in response to changes in the *price level*.

short-run competitive equilibrium The *price* at which the total quantity of a product supplied in the *short run* in a purely competitive *industry* equals the total quantity of the product demanded and that is equal to or greater than *average variable cost*.

short-run supply curve A *supply curve* that shows the quantity of a product a *firm* in a purely competitive *industry* will offer to sell at various *prices* in the *short run;* the portion of the firm's short-run *marginal cost curve* that lies above its *average-variable-cost* curve.

shutdown case The circumstance in which a *firm* would experience a loss greater than its total *fixed cost* if it were to produce any output greater than zero; alternatively, a situation in which a firm would cease to operate when the *price* at which it can sell its product is less than its *average variable cost*.

simple multiplier The *multiplier* in any economy in which government collects no *net taxes*, there are no *imports*, and *investment* is independent of the level of *income;* equal to 1 divided by the *marginal propensity to save*.

simultaneous consumption The same-time derivation of *utility* from some product by a large number of consumers.

simultaneous game In *game theory*, a game in which both parties choose their strategies and execute them at the same time.

single-tax movement The political efforts by followers of Henry George (1839–1897) to impose a single *tax* on the value of land and eliminate all other taxes.

skill transferability The ease with which people can shift their work talents from one job, region, or country to another job, region, or country.

slope of a straight line The ratio of the vertical change (the rise or fall) to the horizontal change (the run) between any two points on a straight line. The slope of an upward-sloping line is positive, reflecting a direct relationship between two variables; the slope of a downward-sloping line is negative, reflecting an inverse relationship between two variables.

Smoot-Hawley Tariff Act Legislation passed in 1930 that established very high *tariffs*. Its objective was to reduce *imports* and stimulate the domestic economy, but it resulted only in retaliatory tariffs by other nations.

social insurance programs Programs that replace the earnings lost when people retire or are temporarily unemployed, that are financed by payroll *taxes*, and that are viewed as earned rights (rather than charity).

socially optimal price The *price* of a product that results in the most efficient allocation of an economy's *resources* and that is equal to the *marginal cost* of the product.

social regulation Regulation in which government is concerned with the conditions under which *goods* and *services* are produced, their physical characteristics, and the impact of their production on society. Differs from *industrial regulation*.

Social Security The social insurance program in the United States financed by federal *payroll taxes* on employers and employees and designed to replace a portion of the earnings lost when workers become disabled, retire, or die.

Social Security trust fund A federal fund that saves excessive Social Security *tax* revenues received in one year to meet Social Security benefit obligations that exceed Social Security tax revenues in some subsequent year.

sole proprietorship An unincorporated *firm* owned and operated by one person.

special-interest effect Any political outcome in which a small group ("special interest") gains substantially at the expense of a much larger number of persons who each individually suffers a small loss.

specialization The use of the *resources* of an individual, a *firm*, a region, or a nation to concentrate production on one or a small number of *goods* and *services*.

speculation The activity of buying or selling with the motive of later reselling or rebuying for *profit*.

SSI (See *Supplemental Security Income*.)

stagflation *Inflation* accompanied by stagnation in the rate of growth of output and an increase in *unemployment* in the economy; simultaneous increases in the *inflation rate* and the *unemployment rate*.

standardized product A product whose buyers are indifferent to the seller from whom they purchase it as long as the *price* charged by all sellers is the same; a product all units of which are identical and thus are perfect substitutes for each other.

Standard Oil case A 1911 antitrust case in which Standard Oil was found guilty of violating the *Sherman Act* by illegally monopolizing the petroleum *industry*. As a remedy the company was divided into several competing *firms*.

start-up firm A new *firm* focused on creating and introducing a particular new product or employing a specific new production or distribution method.

state bank A *commercial bank* authorized by a state government to engage in the business of banking.

statistical discrimination The practice of judging an individual on the basis of the average characteristics of the group to which he or she belongs rather than on his or her own personal characteristics.

status quo The existing state of affairs; in *prospect theory*, the current situation from which gains and losses are calculated.

status quo bias The tendency most people have when making choices to select any option that is presented as the default (*status quo*) option. Explainable by *prospect theory* and *loss aversion*.

sticky prices (See *inflexible prices*.)

stock (corporate) An ownership share in a corporation.

stock options Contracts that enable executives or other key employees to buy shares of their employers' *stock* at fixed, lower *prices* even when the market price subsequently rises.

store of value An *asset* set aside for future use; one of the three functions of *money*.

strategic behavior Self-interested economic actions that take into account the expected reactions of others.

strike The withholding of *labor* services by an organized group of workers (a *labor union*).

structural unemployment *Unemployment* of workers whose skills are not demanded by employers, who lack sufficient skill to obtain employment, or who cannot easily move to locations where jobs are available.

subprime mortgage loans High-*interest-rate* loans to home buyers with above-average credit risk.

subsidy A payment of funds (or *goods* and *services*) by a government, *firm*, or household for which it receives no *good* or *service* in return. When made by a government, it is a *government transfer payment*.

substitute goods Products or *services* that can be used in place of each other. When the *price* of one falls, the *demand* for the other product falls; conversely, when the price of one product rises, the demand for the other product rises.

substitute resources Productive inputs that can be used instead of other inputs in the production process; resources for which an increase in the *price* of one leads to an increase in the demand for the other.

substitution effect (1) A change in the quantity demanded of a *consumer good* that results from a change in its relative expensiveness caused by a change in the good's own *price;* (2) the reduction in the *quantity demanded* of the second of a pair of *substitute resources* that occurs when the price of the first resource falls and causes *firms* that employ both resources to switch to using more of the first resource (whose price has fallen) and less of the second resource (whose price has remained the same).

sunk cost A cost that has been incurred and cannot be recovered.

Supplemental Nutrition Assistance Program (SNAP) A government program that provides food money to low-income recipients by depositing electronic money onto *Electronic Benefit Transfer* (*EBT*) *cards*. Formerly known as the food-stamp program.

Supplemental Security Income (SSI) A federally financed and administered program that provides a uniform nationwide minimum *income* for the aged, blind, and disabled who do not qualify for benefits under *Social Security* in the United States.

supply A schedule or curve that shows the various amounts of a product that producers are willing and able to make available for sale at each of a series of possible *prices* during a specified period of time.

supply curve A curve that illustrates the *supply* for a product by showing how each possible *price* (on the *vertical axis*) is associated with a specific *quantity supplied* (on the *horizontal axis*).

supply factors (in growth) The four determinants of an economy's physical ability to achieve *economic growth* by increasing *potential output* and shifting out the *production possibilities curve*. The four determinants are improvements in technology plus increases in the quantity and quality of natural resources, human resources, and the stock of capital goods.

supply schedule A table of numbers showing the amounts of a *good* or *service* producers are willing and able to make available for sale at each of a series of possible *prices* during a specified period of time.

supply shocks Sudden, unexpected changes in *aggregate supply*.

supply-side economics A view of *macroeconomics* that emphasizes the role of costs and *aggregate supply* in explaining *inflation*, *unemployment*, and *economic growth*.

supply-side market failures Overallocations of *resources* that occur when private supply curves understate the full cost of producing a *good* or *service*.

surplus The amount by which the *quantity supplied* of a product exceeds the *quantity demanded* at a specific (above-equilibrium) *price*.

surplus payment A payment exceeding the minimum payment necessary to ensure the availability of a *resource* in a production process; for example, land rent.

systematic errors Suboptimal choices that (1) are not *rational* because they do not maximize a person's chances of achieving his or her goals and (2) occur routinely, repeatedly, and predictably.

tacit understanding An unspoken, unwritten agreement by an oligopolist to set *prices* and outputs that does not involve outright (or overt) *collusion. Price leadership* is a frequent example.

TANF (See *Temporary Assistance for Needy Families*.)

tariff A *tax* imposed by a nation on an imported good.

taste-for-discrimination model A theory that views discrimination as a preference for which an employer is willing to pay.

tax An involuntary payment of money (or *goods* and *services*) to a government by a *household* or *firm* for which the household or firm receives no good or service directly in return.

tax credit An accounting *credit* that reduces the amount of *taxes* owed to the government on a dollar-for-dollar basis. Some tax credits are refundable, meaning that if the amount of the tax credit exceeds the amount owed in taxes, the taxpayer will receive (be refunded) the difference in cash.

taxes on production and imports A *national income accounting* category that includes such taxes as *sales*, *excise*, business property taxes, and *tariffs* that *firms* treat as costs of producing a product and pass on (in whole or in part) to buyers by charging a higher *price*.

tax incidence The degree to which a *tax* falls on a particular person or group.

tax subsidy A grant in the form of reduced *taxes* through favorable *tax* treatment. For example, employer-paid health insurance is exempt from federal *income taxes* and *payroll taxes*.

Taylor rule A *monetary rule* proposed by economist John Taylor that would stipulate exactly how much the *Federal Reserve System* should change *real interest rates* in response to divergences of *real GDP* from potential GDP and divergences of actual rates of *inflation* from a target rate of inflation.

technological advance (Web chapter) (1) An improvement in the quality of existing products, the invention of entirely new products, or the creation of new or better ways of producing or distributing products. (2) Any improvement in the methods by which resources are combined such that the same quantity of inputs can be made to yield a combination of outputs that is preferred to any combination of outputs that was previously possible.

technology The body of knowledge and techniques that can be used to combine *economic resources* to produce *goods* and *services*.

Temporary Assistance for Needy Families (TANF) A state-administered and partly federally funded program in the United States that provides financial aid to poor families; the basic welfare program for low-income families in the United States; contains time limits and work requirements.

term auction facility A *monetary policy* procedure used by the *Federal Reserve System* during the 2008 financial crisis. *Commercial banks* anonymously bid to obtain loans being made available by the Fed as a way to expand *reserves* in the banking system.

terms of trade The rate at which units of one product can be exchanged for units of another product; the *price* of a *good* or *service;* the amount of one good or service that must be given up to obtain 1 unit of another good or service.

theoretical economics The process of deriving and applying *economic theories* and principles.

theory of human capital The generalization that *wage differentials* are the result of differences in the amount of *human capital investment* and that the *incomes* of lower-paid workers are raised by increasing the amount of such investment.

thrift institution A *savings and loan association*, *mutual savings bank*, or *credit union*.

till money (See *vault cash*.)

time deposit An interest-earning deposit in a *commercial bank* or *thrift institution* that the depositor can withdraw without penalty after the end of a specified period.

time inconsistency The human tendency to systematically misjudge at the present time what will actually end up being desired at a future time.

time preference The human tendency for people, because of impatience, to prefer to spend and consume in the present rather than save and wait to spend and consume in the future; this inclination varies in strength among individuals.

time-value of money The idea that a specific amount of *money* is more valuable to a person the sooner it is received because the money can be placed in a financial account or *investment* and earn *compound interest* over time; the *opportunity cost* of receiving a sum of money later rather than earlier.

token money Bills or coins for which the amount printed on the *currency* bears no relationship to the value of the paper or metal embodied within it; for currency still circulating, *money* for which the face value exceeds the commodity value.

total allowable catch (TAC) The overall limit set by a government or a fisheries commission on the total number of fish or tonnage of fish that fishers collectively can harvest during some particular time period. Used to set the fishing limits for *individual transferable quotas* (*ITQs*).

total cost The sum of *fixed cost* and *variable cost*.

total demand The *demand schedule* or the *demand curve* of all buyers of a *good* or *service;* also called market demand.

total demand for money The sum of the *transactions demand for money* and the *asset demand for money*.

total fertility rate The average number of children per lifetime birthed by a nation's women.

total product (TP) The total output of a particular *good* or *service* produced by a *firm* (or a group of firms or the entire economy).

total revenue (TR) The total number of dollars received by a *firm* (or firms) from the sale of a product; equal to the total expenditures for the product produced by the firm (or firms); equal to the quantity sold (demanded) multiplied by the *price* at which it is sold.

total-revenue test A test to determine *elasticity of demand*. Demand is elastic if *total revenue* moves in the opposite direction from a *price* change; it is inelastic when it moves in the same direction as a price change; and it is of unitary elasticity when it does not change when price changes.

total spending The total amount that buyers of *goods* and *services* spend or plan to spend; also called *aggregate expenditures*.

total supply The *supply schedule* or the *supply curve* of all sellers of a *good* or *service;* also called market supply.

total utility The total amount of satisfaction derived from the consumption of a single product or a combination of products.

Trade Adjustment Assistance Act A U.S. law passed in 2002 that provides cash assistance, education and training benefits, health care subsidies, and *wage* subsidies (for persons age 50 or older) to workers displaced by *imports* or relocations of U.S. *plants* to other countries.

trade balance The export of *goods* (or goods and *services*) of a nation less its imports of goods (or goods and *services*).

trade controls *Tariffs*, *export subsidies*, *import quotas*, and other means a nation may employ to reduce *imports* and expand *exports*.

trade deficit The amount by which a nation's *imports* of *goods* (or goods and *services*) exceed its *exports* of goods (or goods and services).

trademark A word, symbol, or phrase that serves to distinguish a product produced by a particular *firm* from products produced by rival firms. Most countries register and enforce trademarks so that the originator of a particular product is the only firm that may utilize its legally registered trademarks.

trade-off The sacrifice of some or all of one economic goal, *good*, or *service* to achieve some other goal, good, or service.

trade surplus The amount by which a nation's *exports* of *goods* (or goods and *services*) exceed its *imports* of goods (or goods and services).

trading possibilities line A line that shows the different combinations of two products that an economy is able to obtain (consume) when it specializes in the production of one product and trades (exports) it to obtain the other product.

tragedy of the commons The tendency for commonly owned *natural resources* to be overused, neglected, or degraded because their common ownership gives nobody an incentive to maintain or improve them.

transactions demand for money The amount of money people want to hold for use as a *medium of exchange* (to make payments); varies directly with *nominal GDP*.

transfer payment A payment of *money* (or *goods* and *services*) by a government to a *household* or *firm* for which the payer receives no *good* or *service* directly in return.

Troubled Asset Relief Program (TARP) A 2008 federal government program that authorized the U.S. Treasury to loan up to \$700 billion to critical financial institutions and other U.S. *firms* that were in extreme financial trouble and therefore at high risk of failure.

trough The point in a *business cycle* at which business activity has reached a temporary minimum; the point at which a *recession* ends and an *expansion* (recovery) begins. At the trough, the economy experiences substantial *unemployment* and *real GDP* is less than *potential output*.

tying contract A requirement imposed by a seller that a buyer purchase another (or other) of its products as a condition for buying a desired product; a practice forbidden by the *Clayton Act*.

ultimatum game A *behavioral economics* game in which a mutually anonymous pair of players interact to determine how an amount of money is to be split. The first player suggests a division. The second player either accepts that proposal (in which case the split is made accordingly) or rejects it (in which case neither player gets anything).

unanticipated inflation An increase of the *price level* (*inflation*) at a rate greater than expected.

underemployment (Web chapter) A situation in which workers are employed in positions requiring less education and skill than they have.

undistributed corporate profits After-*tax* corporate *profits* not distributed as *dividends* to stockholders; corporate or business *saving;* also called retained earnings.

unemployment The failure to use all available *economic resources* to produce desired *goods* and *services;* the failure of the economy to fully employ its *labor force*.

unemployment compensation (See *unemployment insurance*).

unemployment insurance The social insurance program that in the United States is financed by state *payroll taxes* on employers and makes *income* available to workers who become unemployed and are unable to find jobs.

unemployment rate The percentage of the *labor force* unemployed at any time.

unfulfilled expectations Situations in which *households* and businesses were expecting one thing to happen but instead find that something else has happened; unrealized anticipations or plans relating to future economic conditions and outcomes.

unfunded liability A future government spending commitment (liability) for which the government has not legislated an offsetting revenue source.

uninsurable risk An eventuality for which the frequency or magnitude of potential losses is unpredictable or unknowable. Insurance companies are not willing to sell insurance against such risks.

unintended consequences Unexpected results of government policies. Can be good or bad, but normally refer to unexpected negative outcomes.

union (See *labor union*.)

unionization rate The percentage of a particular population of workers that belongs to *labor unions;* alternatively, the percentage of a population of workers that is represented by one union or another in *collective bargaining*.

union shop A place of employment where the employer may hire either *labor union* members or nonmembers but where nonmembers must become members within a specified period of time or lose their jobs.

unit elasticity *Demand* or *supply* for which the *elasticity coefficient* is equal to 1; means that the percentage change in the *quantity demanded* or *quantity supplied* is equal to the percentage change in *price*.

unit labor cost *Labor* cost per unit of output; total labor cost divided by total output; also equal to the *nominal wage* rate divided by the *average product* of labor.

unit of account A standard unit in which *prices* can be stated and the value of *goods* and *services* can be compared; one of the three functions of *money*.

unlimited wants The insatiable desire of consumers for *goods* and *services* that will give them satisfaction or *utility*.

unplanned changes in inventories Changes in *inventories* that *firms* did not anticipate; changes in inventories that occur because of unexpected increases or decreases of aggregate spending (or of *aggregate expenditures*).

unplanned investment *Actual investment* less *planned investment;* increases or decreases in the *inventories* of *firms* resulting from production greater than sales.

Uruguay Round A 1995 trade agreement (fully implemented in 2005) that established the *World Trade Organization* (*WTO*), liberalized trade in *goods* and *services*, provided added protection to intellectual property (for example, *patents* and *copyrights*), and reduced farm *subsidies*.

user cost The *opportunity cost* of extracting and selling a *nonrenewable natural resource* today rather than waiting to extract and sell the resource in the future; the *present value* of the decline in future revenue that will occur because a nonrenewable natural resource is extracted and sold today rather than being extracted and sold in the future.

U.S. government securities U.S. Treasury bills, notes, and *bonds* used to finance *budget deficits;* the components of the *public debt*.

U.S. Steel case The antitrust action brought by the federal government against the U.S. Steel Corporation in which the courts ruled (in 1920) that only unreasonable restraints of trade were illegal and that size and the possession of monopoly power were not by themselves violations of the *antitrust laws*.

usury laws State laws that specify the maximum legal *interest rate* at which loans can be made.

utility The want-satisfying power of a *good* or *service;* the satisfaction or pleasure a consumer obtains from the consumption of a good or service (or from the consumption of a collection of *goods* and *services*).

utility-maximizing rule The principle that to obtain the greatest *total utility*, a consumer should allocate *money income* so that the last dollar spent on each *good* or *service* yields the same *marginal utility* (MU). For two goods X and Y, with prices P_x and P_y, total utility will be maximized by purchasing the amounts of X and Y such that $MU_x/P_x = MU_y/P_y$ for the last dollar spent on each good.

value added The value of a product sold by a *firm* less the value of the products (materials) purchased and used by the firm to produce that product.

value-added tax A *tax* imposed on the difference between the value of a product sold by a *firm* and the value of the goods purchased from other firms to produce that product; used in several European countries.

value judgment Opinion of what is desirable or undesirable; belief regarding what ought or ought not to be in terms of what is right (or just) or wrong (or unjust).

value of money The quantity of *goods* and *services* for which a unit of *money* (a dollar) can be exchanged; the purchasing power of a unit of money; the reciprocal of the *price index*.

variable cost A cost that increases when the *firm* increases its output and decreases when the firm reduces its output.

VAT (See *value-added tax*.)

vault cash The *currency* a bank has on hand in its vault and cash drawers.

velocity The number of times per year that the average dollar in the *money supply* is spent for *final goods* and final *services; nominal gross domestic product* (*GDP*) divided by the *money supply*.

venture capital (Web chapter) That part of household *saving* used to finance high-risk business enterprises in exchange for *stock* (and thus a share of any *profit* if the enterprises are successful).

vertical axis The "up-down" or "north-south" measurement line on a graph or grid.

vertical integration Any situation in which a group of *plants* that are each engaged in a different stage of the production of a particular *final good* are all owned by a single *firm*.

vertical intercept The point at which a line meets the vertical axis of a graph.

vertical merger The merger of one or more *firms* engaged in different stages of the production of a particular *final good*.

very long run (Web chapter) In microeconomics, a period of time long enough that *technology* can change and *firms* can introduce new products.

vicious circle of poverty (Web chapter) A problem common in some *developing countries* in which their low *per capita incomes* are an obstacle to realizing the levels of *savings* and *investment* needed to achieve rates of growth of output that exceed their rates of population growth.

voice mechanism Communication by workers through their *union* to resolve grievances with an employer.

voluntary export restrictions (VER) Voluntary limitations by countries or *firms* of their exports to a particular foreign nation; undertaken to avoid the enactment of formal trade barriers by the foreign nation.

wage The *price* paid for the use or *services* of *labor* per unit of time (per hour, per day, and so on).

wage differential The difference between the *wage* received by one worker or group of workers and that received by another worker or group of workers.

wage rate (See *wage*.)

wages The *income* that compensates for *labor*; *earnings*.

Wall Street Reform and Consumer Protection Act of 2010 The law that gave authority to the *Federal Reserve System* (the Fed) to regulate all large financial institutions, created an oversight council to look for growing risk to the financial system, established a process for the federal government to sell off the assets of large failing financial institutions, provided federal regulatory oversight of asset-backed *securities*, and created a financial consumer protection bureau within the Fed.

wealth Anything that has value because it produces *income* or could produce income. Wealth is a stock; income is a flow. Assets less liabilities; net worth.

wealth effect The tendency for people to increase their consumption spending when the value of their financial and real *assets* rises and to decrease their consumption spending when the value of those assets falls.

welfare programs (See *public assistance programs*.)

Wheeler-Lea Act The federal law of 1938 that amended the *Federal Trade Commission Act* by prohibiting unfair and deceptive acts or practices of commerce (such as false and misleading advertising and the misrepresentation of products).

will to develop (Web chapter) The mental state of wanting *economic growth* strongly enough to change from old to new ways of doing things.

World Bank (Web chapter) A bank that lends (and guarantees loans) to *developing countries* to assist them in increasing their *capital stock* and thus in achieving *economic growth*.

world price The international market *price* of a *good* or *service*, determined by world demand and supply.

World Trade Organization (WTO) An organization of 159 nations (as of mid-2013) that oversees the provisions of the current world trade agreement, resolves trade disputes stemming from it, and holds forums for further rounds of trade negotiations.

WTO (See *World Trade Organization*.)

X-inefficiency The production of output, whatever its level, at a higher average (and total) cost than is necessary for producing that level of output.

zero interest rate policy (ZIRP) A *monetary policy* in which a *central bank* sets *nominal interest rates* at or near zero percent per year in order to stimulate the economy.

zero lower bound problem The constraint placed on the ability of a *central bank* to stimulate the economy through lower interest rates by the fact that *nominal interest rates* cannot be driven lower than zero (because if interest rates were negative, people would be unwilling to put their money into banks due to the fact that deposit balances would decrease over time due to the negative interest rate.)

zero-sum game In *game theory*, a game in which the gains (+) and losses (−) add up to zero; one party's gain equals the other party's loss.

PHOTO CREDITS

Part and chapter openers: Ingram Publishing.

Chapter 1 Page 2: © Dougal Waters/Getty Images; p. 3: © David Leahy/Getty Images; p. 5: Courtesy of Robbins Recreation, www.4americanrecreation.com, and artist Katherine Robbins; p. 6: © Syracuse Newspapers/The Image Works; p. 11: © Mike Theiler/Reuters/Corbis; p. 15: © Royalty-Free/Corbis; p. 18: © James Leynse/Corbis.

Chapter 2 Page 38: © The McGraw-Hill Companies, Inc./John Flournoy, photographer; p. 42: © Stocktrek Images/Getty Images; p. 46 © AGE Fotostock; p. 47: © Royalty-Free/Corbis.

Chapter 3 Page 64: © Robert W. Ginn/PhotoEdit; p. 65: © Nancy R. Cohen/Getty Images; p. 69: © Olivier Matthys/epa/Corbis

Chapter 4 Page 92: © Colin Young-Wolff/PhotoEdit; p.93: © Getty Images; p. 98: © Digital Vision/Getty Images; p. 103: © Royalty-Free/Corbis.

Chapter 5 Page 114: © Getty Images; p. 116: © Hein von Horsten/Getty Images; p. 119: © Getty Images; p. 123: © Royalty-Free/Corbis; p. 130: © SuperStock.

Chapter 6 Page 137: © Nancy R. Cohen/Getty Images; p. 144: © Punchstock; p. 147: © Tom Prettyman/PhotoEdit .

Chapter 7 Page 155: © Kim Kulish/Corbis; p. 161: © Digital Vision/Getty Images; p. 170: © Ryan McVay/Getty Images.

Chapter 8 Page 176: © Bettmann/Corbis; p. 181: © Stewart Cohen/Index Stock/Photolibrary; p. 186: © Corbis/PunchStock; p. 188: © Manoj Shah/Getty Images.

Chapter 9 Page 201: © BananaStock/JupiterImages; p. 206: © Don Farrall/Getty Images; p. 214: © Getty Images.

Chapter 10 Page 232: © Getty Images; p. 234: © iStockphoto.

Chapter 11 Page 250: © Katrin Thomas/Getty Images; p. 249: © MCT Graphics via Getty Images.

Chapter 12 Page 269: © Getty Images; p. 272: © Anatolii Babii/Alamy.

Chapter 13 Page 287: © Richard Cummins/Corbis; p. 289: © Brand X Pictures; p. 299: © Anatolii Babii/Alamy.

Chapter 13W Page 13W-12: © Don Farrall/Getty Images.

Chapter 14 Page 317: © PRNewsFoto/Diamond Information Center; p. 325: © age fotostock/SuperStock.

Chapter 15 Page 335: © numbeos/Getty Images; p. 345: Reunion des Musees Nationaux/Art Resource, NY. © 2007 Estate of Pablo Picasso/Artiss Rights Society (ARS), New York; p. 348: © Peter Weber/Getty Images.

Chapter 16 Page 369: © Matthew Borkoski Photography/Getty Images; p. 373: © AFP/Getty Images; p. 374: © Royalty-Free/Corbis.

Chapter 17 Page 382: © Digital Vision/Getty Images; p. 388: © Henglein and Steets/Getty Images; p. 400: © Digital Vision/PunchStock; p. 399: © PhotoLink/Getty Images.

Chapter 18 Page 412: © Mark Steinmetz/Amanita Pictures; p. 415: © Steve Cole/Getty Images; p. 422: © PhotoDisc/Getty Images.

Chapter 19 Page 435: © AP Photo/Stephen Lance Dennee; p. 440: © AP Photo/Paul Sakuma.

Chapter 20 Page 451: © Digital Vision/PunchStock; p. 456: © Burke/Triolo Productions/Getty Images; p. 461: © Envision/Corbis.

Chapter 21 Page 472: © PrNewsFoto/Dream Works Home Entertainment; p. 474: © Tom Grill/Corbis; p. 484: © Kelly Ryerson/Getty Images.

Chapter 22 Page 498: © Oleg (RF)/Alamy; p. 502:© BlendImages/Alamy; p. 509: © Comstock Images/PictureQuest; p. 506: © Fuzzbones/Alamy.

Chapter 23 Page 518: © Royalty-Free/Corbis; p. 525: © Getty Images.

Chapter 24 Page 535: © Getty Images; p. 539: © Getty Images; p. 541: © ImagesState Media (John Foxx).

Chapter 25 Page 553: © photo by Lynn Betts, USDA Natural Resources Conservation Service; p. 562: © Getty Images.

Chapter 26 Page 573: © Getty Images; p. 575: © Photodisc Collection/Getty Images; p. 579: © Dynamic Graphics/JupiterImages; p. 586: © Caro/Alamy.

Chapter 27 Page 596: © JGI/Jamie Grill/Getty Images; p. 603: © Royalty-Free/Corbis; p. 604: © Spencer Platt/Staff/Getty Images; p.608: © Jim West/Alamy.

Chapter 28 Page 621: © Digital Vision/Getty Images; p. 626: © Getty Images; p. 631: © Getty Images.

Chapter 29 Page 653: Cover image from The General Theory of Employment, Interest, and Money, by John Maynard Keynes. (Amherst, NY: Prometheus Books, 1997). Copyright © 1997 by Prometheus Books. Reprinted with permission of the publisher.

Chapter 30 Page 673: © Royalty-Free/Corbis; p. 676: © rumal/Alamy.

Chapter 31 Page 702: © Royalty-Free/Corbis.

Chapter 32 Page 713: © Getty Images; p. 724: © J. Scott Applewhite/AP Images.

Chapter 33 Page 742: © Bloomberg via Getty Images.

Chapter 34 Page 760: © Royalty-Free/Corbis; p. 768: © Steve Allen/Getty Images; p. 772: © Kyoungil Jeon/Getty Images.

Chapter 35 Page 791: © AP Photo/Mary Altaffer; p. 793: © Creasource/Corbis.

Chapter 36 Page 814: © Bettmann/Corbis; p. 816: © Thinkstock/Corbis.

Chapter 37 Page 824: © Baron Wolman/Getty Images; p. 829: © Getty Images; p. 832: © Kulka/zefa/Corbis.

Chapter 38 Page 841: © Digital Vision/Getty Images; p. 847: © Bloomberg via Getty Images; p. 852: © Robert W. Ginn/PhotoEdit; p. 860: © Steve Allen/Getty Images.

Chapter 39 Page 882: © PhotoLink/Getty Images.

Chapter 39W Page 39W-16: © Jon Jones/Sygma/Corbis.

COI1 Page COI1-8: © Digital Vision/Getty Images; Page COI1-13: © Robert W. Ginn/PhotoEdit; Page COI1-18: © Getty Images.

INDEX

Note: **Bold** type indicates key terms and definitions; page numbers followed by "n" indicate notes.